

BIRCH, STEWART, KOLASCH & BIRCH, LLP

TERRELL C. BIRCH
RAYMOND C. STEWART
JOSEPH A. KOLASCH
JAMES M. SLATTERY
BERNARD L. SWEENEY*
MICHAEL K. MUTTER
CHARLES GORENSTEIN
GERALD M. MURPHY, JR.
LEONARD R. SVENSSON
EDWARD L. CLARK
DREW D. MEIKLE
RICHARD S. WEINER
JOHN MCKINNEY MUNCY
ROBERT J. KENNEY
DONALD J. DALEY
JOHN W. BAILEY
JOHN A. CASTELLANO, III
DAVID D. YACURA

OF COUNSEL:
HERBERT M. BIRCH (1905-1996)
ELLIOT A. GOLDBERG*
WILLIAM L. GATES*
EDWARD H. VALANCE
RUPERT J. BRADY (RET)*
F. PRINCE BUTLER
FRED S. WHISENHUNT

INTELLECTUAL PROPERTY LAW
8110 GATEHOUSE ROAD
SUITE 500 EAST
FALLS CHURCH, VA 22042-1210
U S A
(703) 205-8000

FAX: (703) 205-8050
(703) 698-8590 (G IV)

e-mail: mailroom@bskb.com
web: <http://www.bskb.com>

CALIFORNIA OFFICE:
COSTA MESA, CALIFORNIA

THOMAS S. AUCHTERLONIE
JAMES T. ELLER, JR.
SCOTT L. LOWE
MARK J. NUEL, PH D
D. RICHARD ANDERSON
PAUL C. LEWIS
MARK W. MILSTEAD*
RICHARD J. GALLAGHER
JAYNE M. SAYDAH*

REG. PATENT AGENTS
FREDERICK R. HANDREN
MARYANNE ARMSTRONG, PH D
MAKI HATSUMI
MIKE S. RYU
CRAIG A. MCROBBIE
GARTH M. DAHLEN, PH D
LAURA C. LUTZ
ROBERT E. GOOZNER, PH D
HYUNG N. SOHN
MATTHEW J. LATTIG
ALAN PEDERSEN-GILES
C. KEITH MONTGOMERY
TIMOTHY R. WYCKOFF
KRISTIL RUPERT, PH D
LARRY J. HUME
HARAY A. SAYADIAN, PH D

*ADMITTED TO A BAR OTHER THAN VA

Date: August 11, 2000

Docket No.: 2750-1096P

BOX PATENT APPLICATION

Assistant Commissioner for Patents
Washington, DC 20231

Sir:

As authorized by the inventor(s), transmitted herewith for filing is a patent application applied for on behalf of the inventor(s) according to the provisions of 37 C.F.R. § 1.41(c), which claims priority under 35 U.S.C. § 119(e) of Provisional Application No. 60/148,684 filed on August 13, 1999

Inventor(s): Nickolai ALEXANDROV, Vyacheslav BROVER

For: SEQUENCE-DETERMINED DNA FRAGMENTS AND CORRESPONDING POLYPEPTIDES ENCODED THEREBY

Enclosed are:

- ☒ A specification consisting of a Description (1048 pages), Table 1 (80 pages), Table 2 (309 pages), Claims (5 pages), schematic (1 page), Abstract (1 page) totaling one-thousand four-hundred and forty-four (1444) pages
- ☐ () sheet(s) of formal drawings
- ☐ Certified copy of Priority Document(s)
- ☒ Executed Declaration in accordance with 37 C.F.R. § 1.64 will follow
- ☒ A statement to establish small entity status under 37 C.F.R. § 1.9 and 37 C.F.R. § 1.27

- ☐ Preliminary Amendment
- ☒ Information Sheet
- ☐ Information Disclosure Statement, PTO-1449 and reference(s)
- ☐ Amend the specification by inserting before the first line the sentence:

--This application claims priority on provisional Application No. filed on , the entire contents of which are hereby incorporated by reference.--

- ☒ Other: Power of Attorney regarding Small Entity Statement, ATCC Deposit receipts PTA-595, PTA-1161, PTA-1411, CD containing Specification

The filing fee has been calculated as shown below:

			LARGE ENTITY	SMALL ENTITY
BASIC FEE			\$690.00	\$345.00
	NUMBER FILED	NUMBER EXTRA	RATE FEE	RATE FEE
TOTAL CLAIMS	50- 20 =	30	X 18 = \$0.00	x 9 = 270
INDEPENDENT CLAIMS	5- 3 =	2	x 78 = \$0.00	x 39 = 78
<input type="checkbox"/> MULTIPLE DEPENDENT CLAIMS PRESENTED			+ \$260.00	+ \$130.00
TOTAL			\$0.00	\$693.00

- ☒ The application transmitted herewith is filed in accordance with 37 C.F.R. § 1.41(c). The undersigned has been authorized by the inventor(s) to file the present application. The original duly executed declaration together with the surcharge will be forwarded in due course.
- ☒ A check in the amount of \$693.00 to cover the filing fee is enclosed.

☐ Please charge Deposit Account No. 02-2448 in the amount of \$0.00. A triplicate copy of this transmittal form is enclosed.

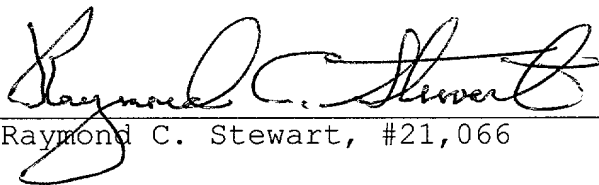
☒ Please send correspondence to:

BIRCH, STEWART, KOLASCH & BIRCH, LLP **or** Customer No. 2292
P.O. Box 747
Falls Church, VA 22040-0747
Telephone: (703) 205-8000

If necessary, the Commissioner is hereby authorized in this, concurrent, and future replies, to charge payment or credit any overpayment to Deposit Account No. 02-2448 for any additional fees required under 37 C.F.R. §§ 1.16 or 1.17; particularly, extension of time fees.

Respectfully submitted,

BIRCH, STEWART, KOLASCH & BIRCH, LLP

By 
Raymond C. Stewart, #21,066

P.O. Box 747
Falls Church, VA 22040-0747
(703) 205-8000

RCS/CAV
2750-1096P

Attachments

(Rev. 04/19/2000)

STATEMENT CLAIMING SMALL ENTITY STATUS
(37 CFR 1.9(f) & 1.27(c)) - SMALL BUSINESS CONCERN

Docket Number: 2750-1096P

Applicant, Patentee, or Identifier: N. ALEXANDROV et al.
Application or Patent No.: NEW Patent Application
Filed or Issued: August 11, 2000
Title: SEQUENCE-DETERMINED DNA FRAGMENTS AND CORRESPONDING POLYPEPTIDES
ENCODED THEREBY

I hereby state that I am

- ☐ the owner of the small business concern identified below:
☒ an official of the small business concern empowered to act on behalf of
the concern identified below:

NAME OF SMALL BUSINESS CONCERN CERES, INC.
ADDRESS OF SMALL BUSINESS CONCERN 3007 Malibu Canyon Road Malibu, CA 90265

I hereby state that the above identified small business concern qualifies as a small business concern as defined in 37 CFR Part 121 for purposes of paying reduced fees to the United States Patent and Trademark Office, in that the number of employees of the concern, including those of its affiliates, does not exceed 500 persons. For purposes of this statement, (1) the number of employees of the business concern is the average over the previous fiscal year of the concern of the persons employed on a full-time, part-time, or temporary basis during each of the pay periods of the fiscal year, and (2) concerns are affiliates of each other when either, directly or indirectly, one concern controls or has the power to control the other, or a third party or parties controls or has the power to control both.

I hereby state that rights under contract or law have been conveyed to and remain with the small business concern identified above with regard to the invention described in:

- ☒ the specification filed herewith with title as listed above.
☐ the application identified above.
☐ the patent identified above.

If the rights held by the above identified small business concern are not exclusive, each individual, concern, or organization having rights in the invention must file separate statements as to their status as small entities, and no rights to the invention are held by any person, other than the inventor, who would not qualify as an independent inventor under 37 CFR 1.9(c) if that person made the invention, or by any concern which would not qualify as a small business concern under 37 CFR 1.9(d), or a nonprofit organization under 37 CFR 1.9(e).

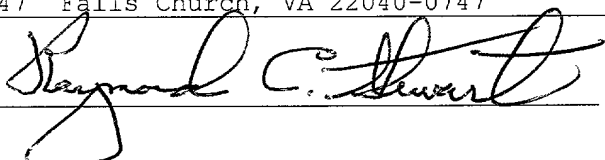
Each person, concern, or organization having any rights in the invention is listed below:

- ☒ no such person, concern, or organization exists.
☐ each such person, concern, or organization is listed below.

Separate statements are required from each named person, concern, or organization having rights to the invention stating their status as small entities. (37 CFR 1.27)

I acknowledge the duty to file, in this application or patent, notification of any change in status resulting in loss of entitlement to small entity status prior to paying, or at the time of paying, the earliest of the issue fee or any maintenance fee due after the date on which status as a small entity is not longer appropriate. (37 CFR 1.28(b))

NAME OF PERSON SIGNING Raymond C. Stewart (Reg. No. 21,066)
TITLE IN ORGANIZATION OF PERSON SIGNING Legal Representative of CERES, INC.
ADDRESS OF PERSON SIGNING Birch, Stewart, Kolasch and Birch, LLP.
P.O. Box 747 Falls Church, VA 22040-0747

SIGNATURE  DATE August 11, 2000

**SEQUENCE-DETERMINED DNA FRAGMENTS AND CORRESPONDING
POLYPEPTIDES ENCODED THEREBY**

This application claims priority under 35 USC §119(e), §119(a-d) and §120 of the following applications, the entire contents of which are hereby incorporated by reference:

Country	Filing Date	Attorney No.	Client No.	Application No.
United States	08/13/99	2750-0532P	80142.002	60/148,684

FIELD OF THE INVENTION

The present invention relates to isolated polynucleotides that represent a complete gene, or a fragment thereof, that is expressed. In addition, the present invention relates to the polypeptide or protein corresponding to the coding sequence of these polynucleotides. The present invention also relates to isolated polynucleotides that represent regulatory regions of genes. The present invention also relates to isolated polynucleotides that represent untranslated regions of genes. The present invention further relates to the use of these isolated polynucleotides and polypeptides and proteins.

DESCRIPTION OF THE RELATED ART

Efforts to map and sequence the genome of a number of organisms are in progress; a few complete genome sequences, for example those of *E. coli* and *Saccharomyces cerevisiae* are known (Blattner et al., *Science* 277:1453 (1997); Goffeau et al., *Science* 274:546 (1996)). The complete genome of a multicellular organism, *C. elegans*, has also been sequenced (See, the *C. elegans* Sequencing Consortium, *Science* 282:2012 (1998)). To date, no complete genome of a plant has been sequenced, nor has a complete cDNA complement of any plant been sequenced.

SUMMARY OF THE INVENTION

The present invention comprises polynucleotides, such as complete cDNA sequences and/or sequences of genomic DNA encompassing complete genes, fragments of genes, and/or regulatory elements of genes and/or regions with other functions and/or intergenic regions, hereinafter collectively referred to as Sequence-Determined DNA Fragments (SDFs), from different plant species, particularly corn, wheat, soybean, rice and *Arabidopsis thaliana*, and other plants and or mutants, variants, fragments or fusions of said SDFs and polypeptides or proteins derived therefrom. In some instances, the SDFs span the entirety of a protein-coding segment. In some instances, the entirety of an mRNA is represented. Other objects of the invention that are also represented by SDFs of the invention are control sequences, such as, but

not limited to, promoters. Complements of any sequence of the invention are also considered part of the invention.

Other objects of the invention are polynucleotides comprising exon sequences, polynucleotides comprising intron sequences, polynucleotides comprising introns together with
5 exons, intron/exon junction sequences, 5' untranslated sequences, and 3' untranslated sequences of the SDFs of the present invention. Polynucleotides representing the joinder of any exons described herein, in any arrangement, for example, to produce a sequence encoding any desirable amino acid sequence are within the scope of the invention.

The present invention also resides in probes useful for isolating and identifying nucleic
10 acids that hybridize to an SDF of the invention. The probes can be of any length, but more typically are 12-2000 nucleotides in length; more typically, 15 to 200 nucleotides long; even more typically, 18 to 100 nucleotides long.

Yet another object of the invention is a method of isolating and/or identifying nucleic acids using the following steps:

- 15 (a) contacting a probe of the instant invention with a polynucleotide sample under conditions that permit hybridization and formation of a polynucleotide duplex; and
(b) detecting and/or isolating the duplex of step (a).

The conditions for hybridization can be from low to moderate to high stringency conditions. The sample can include a polynucleotide having a sequence unique in a plant
20 genome. Probes and methods of the invention are useful, for example, without limitation, for mapping of genetic traits and/or for positional cloning of a desired fragment of genomic DNA.

Probes and methods of the invention can also be used for detecting alternatively spliced messages within a species. Probes and methods of the invention can further be used to detect or isolate related genes in other plant species using genomic DNA (gDNA) and/or cDNA libraries.
25 In some instances, especially when longer probes and low to moderate stringency hybridization conditions are used; the probe will hybridize to a plurality of cDNA and/or gDNA sequences of a plant. This approach is useful for isolating representatives of gene families which are identifiable by possession of a common functional domain in the gene product or which have common cis-acting regulatory sequences. This approach is also useful for identifying
30 orthologous genes from other organisms.

The present invention also resides in constructs for modulating the expression of the genes comprised of all or a fragment of an SDF. The constructs comprise all or a fragment of the expressed SDF, or of a complementary sequence. Examples of constructs include

ribozymes comprising RNA encoded by an SDF or by a sequence complementary thereto, antisense constructs, constructs comprising coding regions or parts thereof, constructs comprising promoters, introns, untranslated regions, scaffold attachment regions, methylating regions, enhancing or reducing regions, DNA and chromatin conformation modifying sequences, etc. Such constructs can be constructed using viral, plasmid, bacterial artificial chromosomes (BACs), plasmid artificial chromosomes (PACs), autonomous plant plasmids, plant artificial chromosomes or other types of vectors and exist in the plant as autonomous replicating sequences or as DNA integrated into the genome. When inserted into a host cell the construct is, preferably, functionally integrated with, or operatively linked to, a heterologous polynucleotide. For instance, a coding region from an SDF might be operably linked to a promoter that is functional in a plant.

The present invention also resides in host cells, including bacterial or yeast cells or plant cells, and plants that harbor constructs such as described above. Another aspect of the invention relates to methods for modulating expression of specific genes in plants by expression of the coding sequence of the constructs, by regulation of expression of one or more endogenous genes in a plant or by suppression of expression of the polynucleotides of the invention in a plant. Methods of modulation of gene expression include without limitation (1) inserting into a host cell additional copies of a polynucleotide comprising a coding sequence; (2) modulating an endogenous promoter in a host cell; (3) inserting antisense or ribozyme constructs into a host cell and (4) inserting into a host cell a polynucleotide comprising a sequence encoding a variant, fragment, or fusion of the native polypeptides of the instant invention.

BRIEF DESCRIPTION OF THE TABLES

The sequences of exemplary SDFs and polypeptides corresponding to the coding sequences of the instant invention are described in Table 1 and Table 2. Table 1 refers to a number of "Maximum Length Sequences" or "MLS." Each MLS corresponds to the longest cDNA obtained, either by cloning or by the prediction from genomic sequence. The sequence of the MLS is the cDNA sequence as described in the Av subsection of Table 1.

Table 1 includes the following information relating to each MLS:

- I. cDNA Sequence
 - A. 5' UTR

- B. Coding Sequence
 - C. 3' UTR
- II. Genomic Sequence
 - A. Exons
 - B. Introns
 - C. Promoters
- III. Link of cDNA Sequences to Clone IDs
- IV. Multiple Transcription Start Sites
- V. Polypeptide Sequences
 - A. Signal Peptide
 - B. Domains
 - C. Related Polypeptides
- VI. Related Polynucleotide Sequences

I. cDNA SEQUENCE

Table 1 indicates which sequence in Table 2 represents the sequence of each MLS. The MLS sequence can comprise 5' and 3' UTR as well as coding sequences. In addition, specific cDNA clone numbers also are included in Table 1 when the MLS sequence relates to a specific cDNA clone.

A. 5' UTR

The location of the 5' UTR can be determined by comparing the most 5' MLS sequence with the corresponding genomic sequence as indicated in Table 1. The sequence that matches, beginning at any of the transcriptional start sites and ending at the last nucleotide before any of the translational start sites corresponds to the 5' UTR.

B. Coding Region

The coding region is the sequence in any open reading frame found in the MLS. Coding regions of interest are indicated in the PolyP SEQ subsection Table 1.

C. 3' UTR

The location of the 3' UTR can be determined by comparing the most 3' MLS sequence with the corresponding genomic sequence as indicated in Table 1. The sequence that matches, beginning at the translational stop site and ending at the last nucleotide of the MLS corresponds to the 3' UTR.

5

II. GENOMIC SEQUENCE

Further, Table 1 indicates the specific "gi" number of the genomic sequence if the sequence resides in a public databank. For each genomic sequence, Table 1 indicates which regions are included in the MLS. These regions can include the 5' and 3' UTRs as well as the coding sequence of the MLS. See, for example, the scheme below:

10

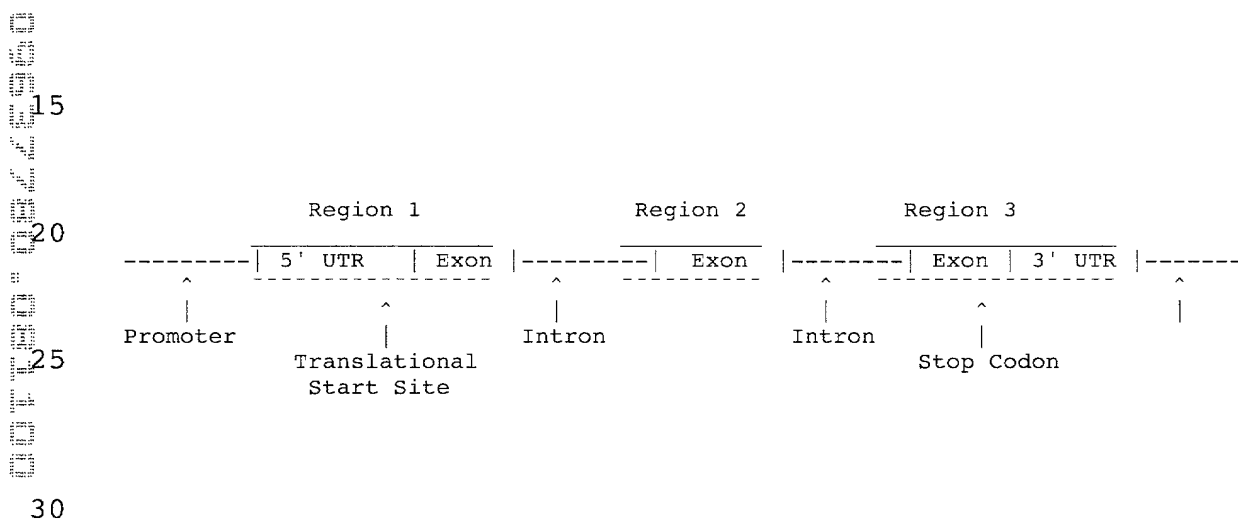


Table 1 reports the first and last base of each region that are included in an MLS sequence. An example is shown below:

35

gi No. 47000:

37102 ... 37497

37593 ... 37925

The numbers indicate that the MLS contains the following sequences from two regions of gi No. 47000; a first region including bases 37102-37497, and a second region including bases 37593-37925.

40

A. EXON SEQUENCES

The location of the exons can be determined by comparing the sequence of the regions from the genomic sequences with the corresponding MLS sequence as indicated by Table 1.

i. INITIAL EXON

To determine the location of the initial exon, information from the

- (1) polypeptide sequence section;
- (2) cDNA polynucleotide section; and
- (3) the genomic sequence section

of Table 1 is used. First, the polypeptide section will indicate where the translational start site is located in the MLS sequence. The MLS sequence can be matched to the genomic sequence that corresponds to the MLS. Based on the match between the MLS and corresponding genomic sequences, the location of the translational start site can be determined in one of the regions of the genomic sequence. The location of this translational start site is the start of the first exon.

Generally, the last base of the exon of the corresponding genomic region, in which the translational start site was located, will represent the end of the initial exon. In some cases, the initial exon will end with a stop codon, when the initial exon is the only exon.

In the case when sequences representing the MLS are in the positive strand of the corresponding genomic sequence, the last base will be a larger number than the first base. When the sequences representing the MLS are in the negative strand of the corresponding genomic sequence, then the last base will be a smaller number than the first base.

ii. INTERNAL EXONS

Except for the regions that comprise the 5' and 3' UTRs, initial exon, and terminal exon, the remaining genomic regions that match the MLS sequence are the internal exons. Specifically, the bases defining the boundaries of the remaining regions also define the intron/exon junctions of the internal exons.

iii. TERMINAL EXON

As with the initial exon, the location of the terminal exon is determined with information from the

- (1) polypeptide sequence section;

- (2) cDNA polynucleotide section; and
- (3) the genomic sequence section

of Table 1. The polypeptide section will indicate where the stop codon is located in the MLS sequence. The MLS sequence can be matched to the corresponding genomic sequence.

Based on the match between MLS and corresponding genomic sequences, the location of the stop codon can be determined in one of the regions of the genomic sequence. The location of this stop codon is the end of the terminal exon. Generally, the first base of the exon of the corresponding genomic region that matches the cDNA sequence, in which the stop codon was located, will represent the beginning of the terminal exon. In some cases, the translational start site will represent the start of the terminal exon, which will be the only exon.

In the case when the MLS sequences are in the positive strand of the corresponding genomic sequence, the last base will be a larger number than the first base. When the MLS sequences are in the negative strand of the corresponding genomic sequence, then the last base will be a smaller number than the first base.

B. INTRON SEQUENCES

In addition, the introns corresponding to the MLS are defined by identifying the genomic sequence located between the regions where the genomic sequence comprises exons. Thus, introns are defined as starting one base downstream of a genomic region comprising an exon, and end one base upstream from a genomic region comprising an exon.

C. PROMOTER SEQUENCES

As indicated below, promoter sequences corresponding to the MLS are defined as sequences upstream of the first exon; more usually, as sequences upstream of the first of multiple transcription start sites; even more usually as sequences about 2,000 nucleotides upstream of the first of multiple transcription start sites.

III. LINK of cDNA SEQUENCES to CLONE IDs

As noted above, Table 1 identifies the cDNA clone(s) that relate to each MLS. The MLS sequence can be longer than the sequences included in the cDNA clones. In such a case, Table 1 indicates the region of the MLS that is included in the clone. If either the 5' or 3' termini of the cDNA clone sequence is the same as the MLS sequence, no mention will be made.

IV. Multiple Transcription Start Sites

Initiation of transcription can occur at a number of sites of the gene. Table 1 indicates the possible multiple transcription sites for each gene. In Table 1, the location of the transcription start sites can be either a positive or negative number.

The positions indicated by positive numbers refer to the transcription start sites as located in the MLS sequence. The negative numbers indicate the transcription start site within the genomic sequence that corresponds to the MLS.

To determine the location of the transcription start sites with the negative numbers, the MLS sequence is aligned with the corresponding genomic sequence. In the instances when a public genomic sequence is referenced, the relevant corresponding genomic sequence can be found by direct reference to the nucleotide sequence indicated by the “gi” number shown in the public genomic DNA section of Table 1. When the position is a negative number, the transcription start site is located in the corresponding genomic sequence upstream of the base that matches the beginning of the MLS sequence in the alignment. The negative number is relative to the first base of the MLS sequence which matches the genomic sequence corresponding to the relevant “gi” number.

In the instances when no public genomic DNA is referenced, the relevant nucleotide sequence for alignment is the nucleotide sequence associated with the amino acid sequence designated by “gi” number of the later PolyP SEQ subsection.

V. Polypeptide Sequences

The PolyP SEQ subsection lists SEQ ID NOs and Ceres SEQ ID NO for polypeptide sequences corresponding to the coding sequence of the MLS sequence and the location of the translational start site with the coding sequence of the MLS sequence.

The MLS sequence can have multiple translational start sites and can be capable of producing more than one polypeptide sequence.

A. Signal Peptide

Table 1 also indicates in subsection (B) the cleavage site of the putative signal peptide of the polypeptide corresponding to the coding sequence of the MLS sequence. Typically, signal peptide coding sequences comprise a sequence encoding the first residue of the polypeptide to the cleavage site residue.

B. Domains

Subsection (C) provides information regarding identified domains (where present) within the polypeptide and (where present) a name for the polypeptide domain.

5

C. Related Polypeptides

Subsection (Dp) provides (where present) information concerning amino acid sequences that are found to be related and have some percentage of sequence identity to the polypeptide sequences of Table 1 and Table 2. These related sequences are identified by a “gi” number.

10

VI. Related Polynucleotide Sequences

Subsection (Dn) provides polynucleotide sequences (where present) that are related to and have some percentage of sequence identity to the MLS or corresponding genomic sequence.

15

Abbreviation	Description
Max Len. Seq.	Maximum Length Sequence
rel to	Related to
Clone Ids	Clone ID numbers
Pub gDNA	Public Genomic DNA
gi No.	gi number
Gen. seq. in cDNA	Genomic Sequence in cDNA (Each region for a single gene prediction is listed on a separate line. In the case of multiple gene predictions, the group of regions relating to a single prediction are separated by a blank line)
(Ac) cDNA SEQ	cDNA sequence
- Pat. Appln. SEQ ID NO	Patent Application SEQ ID NO:
- Ceres SEQ ID NO: 1673877	Ceres SEQ ID NO:
- SEQ # w. TSS	Location within the cDNA sequence, SEQ ID NO:, of Transcription Start Sites which are listed below
- Clone ID #: # -> #	Clone ID comprises bases # to # of the cDNA Sequence
PolyP SEQ	Polypeptide Sequence
- Pat. Appln. SEQ ID NO:	Patent Application SEQ ID NO:
- Ceres SEQ ID NO	Ceres SEQ ID NO:
- Loc. SEQ ID NO: @ nt.	Location of translational start site in cDNA of

Abbreviation	Description
	SEQ ID NO: at nucleotide number
(C) Pred. PP Nom. & Annot.	Nomination and Annotation of Domains within Predicted Polypeptide(s)
- (Title)	Name of Domain
- Loc. SEQ ID NO #: # -> # aa.	Location of the domain within the polypeptide of SEQ ID NO: from # to # amino acid residues.
(Dp) Rel. AA SEQ	Related Amino Acid Sequences
- Align. NO	Alignment number
- gi No	Gi number
- Desp.	Description
- % Idnt.	Percent identity
- Align. Len.	Alignment Length
- Loc. SEQ ID NO: # -> # aa	Location within SEQ ID NO: from # to # amino acid residue.

DETAILED DESCRIPTION OF THE INVENTION

The invention relates to (I) polynucleotides and methods of use thereof, such as

- IA. Probes, Primers and Substrates;
- IB. Methods of Detection and Isolation;
 - B.1. Hybridization;
 - B.2. Methods of Mapping;
 - B.3. Southern Blotting;
 - B.4. Isolating cDNA from Related Organisms;
 - B.5. Isolating and/or Identifying Orthologous Genes
- IC. Methods of Inhibiting Gene Expression
 - C.1. Antisense
 - C.2. Ribozyme Constructs;
 - C.3. Chimeraplasts;
 - C.4. Co-Suppression;
 - C.5. Transcriptional Silencing
 - C.6. Other Methods to Inhibit Gene Expression
- ID. Methods of Functional Analysis;
- IE. Promoter Sequences and Their Use;

- IF. UTRs and/or Intron Sequences and Their Use; and
- IG. Coding Sequences and Their Use.

The invention also relates to (II) polypeptides and proteins and methods of use thereof,
5 such as IIA. Native Polypeptides and Proteins

A.1 Antibodies

A.2 In Vitro Applications

IIB. Polypeptide Variants, Fragments and Fusions

B.1 Variants

B.2 Fragments

B.3 Fusions

The invention also includes (III) methods of modulating polypeptide production, such as

IIIA. Suppression

A.1 Antisense

A.2 Ribozymes

A.3 Co-suppression

A.4 Insertion of Sequences into the Gene to be Modulated

A.5 Promoter Modulation

A.6 Expression of Genes containing Dominant-Negative Mutations

IIIB. Enhanced Expression

B.1 Insertion of an Exogenous Gene

B.2 Promoter Modulation

The invention further concerns (IV) gene constructs and vector construction, such as

IVA. Coding Sequences

IVB. Promoters

IVC. Signal Peptides

The invention still further relates to

V Transformation Techniques

Definitions

Allelic variant An “allelic variant” is an alternative form of the same SDF, which resides at the same chromosomal locus in the organism. Allelic variations can occur in any portion of the gene sequence, including regulatory regions. Allelic variants can arise by normal genetic variation in a population. Allelic variants can also be produced by genetic engineering methods. An allelic variant can be one that is found in a naturally occurring plant, including a cultivar or ecotype. An allelic variant may or may not give rise to a phenotypic change, and may or may not be expressed. An allele can result in a detectable change in the phenotype of the trait represented by the locus. A phenotypically silent allele can give rise to a product.

Alternatively spliced messages Within the context of the current invention, “alternatively spliced messages” refers to mature mRNAs originating from a single gene with variations in the number and/or identity of exons, introns and/or intron-exon junctions.

Chimeric The term “chimeric” is used to describe genes, as defined supra, or constructs wherein at least two of the elements of the gene or construct, such as the promoter and the coding sequence and/or other regulatory sequences and/or filler sequences and/or complements thereof, are heterologous to each other.

Constitutive Promoter: Promoters referred to herein as “constitutive promoters” actively promote transcription under most, but not necessarily all, environmental conditions and states of development or cell differentiation. Examples of constitutive promoters include the cauliflower mosaic virus (CaMV) 35S transcript initiation region and the 1’ or 2’ promoter derived from T-DNA of *Agrobacterium tumefaciens*, and other transcription initiation regions from various plant genes, such as the maize ubiquitin-1 promoter, known to those of skill.

Coordinately Expressed: The term “coordinately expressed,” as used in the current invention, refers to genes that are expressed at the same or a similar time and/or stage and/or under the same or similar environmental conditions.

5 Domain: Domains are fingerprints or signatures that can be used to characterize protein families and/or parts of proteins. Such fingerprints or signatures can comprise conserved (1) primary sequence, (2) secondary structure, and/or (3) three-dimensional conformation. Generally, each domain has been associated with either a family of proteins or motifs. Typically, these families and/or motifs have been correlated with specific *in-vitro* and/or *in-vivo* activities. A domain can be any length, including the entirety of the sequence of a protein. Detailed descriptions of the domains, associated families and motifs, and correlated activities of the polypeptides of the instant invention are described below. Usually, the polypeptides with designated domain(s) can exhibit at least one activity that is exhibited by any polypeptide that comprises the same domain(s).

15 Endogenous The term “endogenous,” within the context of the current invention refers to any polynucleotide, polypeptide or protein sequence which is a natural part of a cell or organisms regenerated from said cell.

20 Exogenous “Exogenous,” as referred to within, is any polynucleotide, polypeptide or protein sequence, whether chimeric or not, that is initially or subsequently introduced into the genome of an individual host cell or the organism regenerated from said host cell by any means other than by a sexual cross. Examples of means by which this can be accomplished are described below, and include *Agrobacterium*-mediated transformation (of dicots - *e.g.* Salomon et al. *EMBO J.* 3:141 (1984); Herrera-Estrella et al. *EMBO J.* 2:987 (1983); of monocots, representative papers are those by Escudero et al., *Plant J.* 10:355 (1996), Ishida et al., *Nature Biotechnology* 14:745 (1996), May et al., *Bio/Technology* 13:486 (1995)), biolistic methods (Armaleo et al., *Current Genetics* 17:97 (1990)), electroporation, *in planta* techniques, and the like. Such a plant containing the exogenous nucleic acid is referred to here as a T₀ for the primary transgenic plant and T₁ for the first generation. The term “exogenous” as used herein is also intended to encompass inserting a naturally found element into a non-naturally found location.

Filler sequence: As used herein, “filler sequence” refers to any nucleotide sequence that is inserted into DNA construct to evoke a particular spacing between particular components such as a promoter and a coding region and may provide an additional attribute such as a restriction enzyme site.

Gene: The term “gene,” as used in the context of the current invention, encompasses all regulatory and coding sequence contiguously associated with a single hereditary unit with a genetic function (see SCHEMATIC 1). Genes can include non-coding sequences that modulate the genetic function that include, but are not limited to, those that specify polyadenylation, transcriptional regulation, DNA conformation, chromatin conformation, extent and position of base methylation and binding sites of proteins that control all of these. Genes comprised of “exons” (coding sequences), which may be interrupted by “introns” (non-coding sequences), encode proteins. A gene’s genetic function may require only RNA expression or protein production, or may only require binding of proteins and/or nucleic acids without associated expression. In certain cases, genes adjacent to one another may share sequence in such a way that one gene will overlap the other. A gene can be found within the genome of an organism, artificial chromosome, plasmid, vector, etc., or as a separate isolated entity.

Gene Family: “Gene family” is used in the current invention to describe a group of functionally related genes, each of which encodes a separate protein.

Heterologous sequences: “Heterologous sequences” are those that are not operatively linked or are not contiguous to each other in nature. For example, a promoter from corn is considered heterologous to an *Arabidopsis* coding region sequence. Also, a promoter from a gene encoding a growth factor from corn is considered heterologous to a sequence encoding the corn receptor for the growth factor. Regulatory element sequences, such as UTRs or 3’ end termination sequences that do not originate in nature from the same gene as the coding sequence originates from, are considered heterologous to said coding sequence. Elements operatively linked in nature and contiguous to each other are not heterologous to each other. On the other hand, these same elements remain operatively linked but become heterologous if other filler sequence is placed between them. Thus, the promoter and coding sequences of a corn gene

expressing an amino acid transporter are not heterologous to each other, but the promoter and coding sequence of a corn gene operatively linked in a novel manner are heterologous.

Homologous gene In the current invention, “homologous gene” refers to a gene that shares sequence similarity with the gene of interest. This similarity may be in only a fragment of the sequence and often represents a functional domain such as, examples including without limitation a DNA binding domain, a domain with tyrosine kinase activity, or the like. The functional activities of homologous genes are not necessarily the same.

Inducible Promoter An “inducible promoter” in the context of the current invention refers to a promoter which is regulated under certain conditions, such as light, chemical concentration, protein concentration, conditions in an organism, cell, or organelle, etc. A typical example of an inducible promoter, which can be utilized with the polynucleotides of the present invention, is PARSK1, the promoter from the *Arabidopsis* gene encoding a serine-threonine kinase enzyme, and which promoter is induced by dehydration, abscissic acid and sodium chloride (Wang and Goodman, *Plant J.* 8:37 (1995)) Examples of environmental conditions that may affect transcription by inducible promoters include anaerobic conditions, elevated temperature, or the presence of light.

Intergenic region “Intergenic region,” as used in the current invention, refers to nucleotide sequence occurring in the genome that separates adjacent genes.

Mutant gene In the current invention, “mutant” refers to a heritable change in DNA sequence at a specific location. Mutants of the current invention may or may not have an associated identifiable function when the mutant gene is transcribed.

Orthologous Gene In the current invention “orthologous gene” refers to a second gene that encodes a gene product that performs a similar function as the product of a first gene. The orthologous gene may also have a degree of sequence similarity to the first gene. The orthologous gene may encode a polypeptide that exhibits a degree of sequence similarity to a polypeptide corresponding to a first gene. The sequence similarity can be found within a

functional domain or along the entire length of the coding sequence of the genes and/or their corresponding polypeptides.

Percentage of sequence identity "Percentage of sequence identity," as used herein, is
5 determined by comparing two optimally aligned sequences over a comparison window, where
the fragment of the polynucleotide or amino acid sequence in the comparison window may
comprise additions or deletions (e.g., gaps or overhangs) as compared to the reference sequence
(which does not comprise additions or deletions) for optimal alignment of the two sequences.
The percentage is calculated by determining the number of positions at which the identical
10 nucleic acid base or amino acid residue occurs in both sequences to yield the number of matched
positions, dividing the number of matched positions by the total number of positions in the
window of comparison and multiplying the result by 100 to yield the percentage of sequence
identity. Optimal alignment of sequences for comparison may be conducted by the local
homology algorithm of Smith and Waterman *Add. APL. Math.* 2:482 (1981), by the homology
15 alignment algorithm of Needleman and Wunsch *J. Mol. Biol.* 48:443 (1970), by the search for
similarity method of Pearson and Lipman *Proc. Natl. Acad. Sci. (USA)* 85: 2444 (1988), by
computerized implementations of these algorithms (GAP, BESTFIT, BLAST, PASTA, and
TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group (GCG), 575
Science Dr., Madison, WI), or by inspection. Given that two sequences have been identified for
20 comparison, GAP and BESTFIT are preferably employed to determine their optimal alignment.
Typically, the default values of 5.00 for gap weight and 0.30 for gap weight length are used.
The term "substantial sequence identity" between polynucleotide or polypeptide sequences
refers to polynucleotide or polypeptide comprising a sequence that has at least 80% sequence
identity, preferably at least 85%, more preferably at least 90% and most preferably at least 95%,
25 even more preferably, at least 96%, 97%, 98% or 99% sequence identity compared to a
reference sequence using the programs.

Plant Promoter A "plant promoter" is a promoter capable of initiating transcription in
plant cells and can drive or facilitate transcription of a fragment of the SDF of the instant
30 invention or a coding sequence of the SDF of the instant invention. Such promoters need not
be of plant origin. For example, promoters derived from plant viruses, such as the CaMV35S
promoter or from *Agrobacterium tumefaciens* such as the T-DNA promoters, can be plant

promoters. A typical example of a plant promoter of plant origin is the maize ubiquitin-1 (ubi-1) promoter known to those of skill.

5

Promoter: The term "promoter," as used herein, refers to a region of sequence determinants located upstream from the start of transcription of a gene and which are involved in recognition and binding of RNA polymerase and other proteins to initiate and modulate transcription. A basal promoter is the minimal sequence necessary for assembly of a transcription complex required for transcription initiation. Basal promoters frequently include a "TATA box" element usually located between 15 and 35 nucleotides upstream from the site of initiation of transcription. Basal promoters also sometimes include a "CCAAT box" element (typically a sequence CCAAT) and/or a GGGCG sequence, usually located between 40 and 200 nucleotides, preferably 60 to 120 nucleotides, upstream from the start site of transcription.

10

15
20

Public sequence: The term "public sequence," as used in the context of the instant application, refers to any sequence that has been deposited in a publicly accessible database. This term encompasses both amino acid and nucleotide sequences. Such sequences are publicly accessible, for example, on the BLAST databases on the NCBI FTP web site (accessible at ncbi.nlm.gov/blast). The database at the NCBI GTP site utilizes "gi" numbers assigned by NCBI as a unique identifier for each sequence in the databases, thereby providing a non-redundant database for sequence from various databases, including GenBank, EMBL, DDBJ, (DNA Database of Japan) and PDB (Brookhaven Protein Data Bank).

25

Regulatory Sequence The term "regulatory sequence," as used in the current invention, refers to any nucleotide sequence that influences transcription or translation initiation and rate, and stability and/or mobility of the transcript or polypeptide product. Regulatory sequences include, but are not limited to, promoters, promoter control elements, protein binding sequences, 5' and 3' UTRs, transcriptional start site, termination sequence, polyadenylation sequence, introns, certain sequences within a coding sequence, etc.

30

Related Sequences: “Related sequences” refer to either a polypeptide or a nucleotide sequence that exhibits some degree of sequence similarity with a sequence described by Table 1 and Table 2.

5 Scaffold Attachment Region (SAR) As used herein, “scaffold attachment region” is a DNA sequence that anchors chromatin to the nuclear matrix or scaffold to generate loop domains that can have either a transcriptionally active or inactive structure (Spiker and Thompson (1996) *Plant Physiol.* 110: 15-21).

10 Sequence-determined DNA fragments (SDFs) “Sequence-determined DNA fragments” as used in the current invention are isolated sequences of genes, fragments of genes, intergenic regions or contiguous DNA from plant genomic DNA or cDNA or RNA the sequence of which has been determined.

15 Signal Peptide A “signal peptide” as used in the current invention is an amino acid sequence that targets the protein for secretion, for transport to an intracellular compartment or organelle or for incorporation into a membrane. Signal peptides are indicated in the tables and a more detailed description located below.

20 Specific Promoter In the context of the current invention, “specific promoters” refers to a subset of inducible promoters that have a high preference for being induced in a specific tissue or cell and/or at a specific time during development of an organism. By “high preference” is meant at least 3-fold, preferably 5-fold, more preferably at least 10-fold still more preferably at least 20-fold, 50-fold or 100-fold increase in transcription in the desired
25 tissue over the transcription in any other tissue. Typical examples of temporal and/or tissue specific promoters of plant origin that can be used with the polynucleotides of the present invention, are: PTA29, a promoter which is capable of driving gene transcription specifically in tapetum and only during anther development (Koltonow et al., *Plant Cell* 2:1201 (1990); RCc2 and RCc3, promoters that direct root-specific gene transcription in rice (Xu et al., *Plant Mol. Biol.* 27:237 (1995); TobRB27, a root-specific promoter from tobacco (Yamamoto et al., *Plant Cell* 3:371 (1991)). Examples of tissue-specific promoters under developmental control include
30 promoters that initiate transcription only in certain tissues or organs, such as root, ovule, fruit,

seeds, or flowers. Other suitable promoters include those from genes encoding storage proteins or the lipid body membrane protein, oleosin. A few root-specific promoters are noted above.

Stringency "Stringency" as used herein is a function of probe length, probe composition (G + C content), and salt concentration, organic solvent concentration, and temperature of hybridization or wash conditions. Stringency is typically compared by the parameter T_m , which is the temperature at which 50% of the complementary molecules in the hybridization are hybridized, in terms of a temperature differential from T_m . High stringency conditions are those providing a condition of $T_m - 5^\circ\text{C}$ to $T_m - 10^\circ\text{C}$. Medium or moderate stringency conditions are those providing $T_m - 20^\circ\text{C}$ to $T_m - 29^\circ\text{C}$. Low stringency conditions are those providing a condition of $T_m - 40^\circ\text{C}$ to $T_m - 48^\circ\text{C}$. The relationship of hybridization conditions to T_m (in $^\circ\text{C}$) is expressed in the mathematical equation

$$T_m = 81.5 - 16.6(\log_{10}[\text{Na}^+]) + 0.41(\%G+C) - (600/N) \quad (1)$$

where N is the length of the probe. This equation works well for probes 14 to 70 nucleotides in length that are identical to the target sequence. The equation below for T_m of DNA-DNA hybrids is useful for probes in the range of 50 to greater than 500 nucleotides, and for conditions that include an organic solvent (formamide).

$$T_m = 81.5 + 16.6 \log \{[\text{Na}^+]/(1 + 0.7[\text{Na}^+])\} + 0.41(\%G+C) - 500/L - 0.63(\%\text{formamide}) \quad (2)$$

where L is the length of the probe in the hybrid. (P. Tijessen, "Hybridization with Nucleic Acid Probes" in Laboratory Techniques in Biochemistry and Molecular Biology, P.C. van der Vliet, ed., c. 1993 by Elsevier, Amsterdam.) The T_m of equation (2) is affected by the nature of the hybrid; for DNA-RNA hybrids T_m is 10-15 $^\circ\text{C}$ higher than calculated, for RNA-RNA hybrids T_m is 20-25 $^\circ\text{C}$ higher. Because the T_m decreases about 1 $^\circ\text{C}$ for each 1% decrease in homology when a long probe is used (Bonner et al., *J. Mol. Biol.* 81:123 (1973)), stringency conditions can be adjusted to favor detection of identical genes or related family members.

Equation (2) is derived assuming equilibrium and therefore, hybridizations according to the present invention are most preferably performed under conditions of probe excess and for sufficient time to achieve equilibrium. The time required to reach equilibrium can be

shortened by inclusion of a hybridization accelerator such as dextran sulfate or another high volume polymer in the hybridization buffer.

Stringency can be controlled during the hybridization reaction or after hybridization has occurred by altering the salt and temperature conditions of the wash solutions used. The formulas shown above are equally valid when used to compute the stringency of a wash solution. Preferred wash solution stringencies lie within the ranges stated above; high stringency is 5-8°C below T_m , medium or moderate stringency is 26-29°C below T_m and low stringency is 45-48°C below T_m .

Substantially free of A composition containing A is “substantially free of “ B when at least 85% by weight of the total A+B in the composition is A. Preferably, A comprises at least about 90% by weight of the total of A+B in the composition, more preferably at least about 95% or even 99% by weight. For example, a plant gene or DNA sequence can be considered substantially free of other plant genes or DNA sequences.

Translational start site In the context of the current invention, a “translational start site” is usually an ATG in the cDNA transcript, more usually the first ATG. A single cDNA, however, may have multiple translational start sites.

Transcription start site “Transcription start site” is used in the current invention to describe the point at which transcription is initiated. This point is typically located about 25 nucleotides downstream from a TFIID binding site, such as a TATA box. Transcription can initiate at one or more sites within the gene, and a single gene may have multiple transcriptional start sites, some of which may be specific for transcription in a particular cell-type or tissue.

Untranslated region (UTR) A “UTR” is any contiguous series of nucleotide bases that is transcribed, but is not translated. These untranslated regions may be associated with particular functions such as increasing mRNA message stability. Examples of UTRs include, but are not limited to polyadenylation signals, terminations sequences, sequences located between the transcriptional start site and the first exon (5' UTR) and sequences located between the last exon and the end of the mRNA (3' UTR).

Variant: The term “variant” is used herein to denote a polypeptide or protein or polynucleotide molecule that differs from others of its kind in some way. For example, polypeptide and protein variants can consist of changes in amino acid sequence and/or charge and/or post-translational modifications (such as glycosylation, etc).

5

DETAILED DESCRIPTION OF THE INVENTION

I. Polynucleotides

Exemplified SDFs of the invention represent fragments of the genome of corn, wheat, rice, soybean or *Arabidopsis* and/or represent mRNA expressed from that genome. The isolated nucleic acid of the invention also encompasses corresponding fragments of the genome and/or cDNA complement of other organisms as described in detail below.

Polynucleotides of the invention can be isolated from polynucleotide libraries using primers comprising sequence similar to those described by Table 1 and Table 2. See, for example, the methods described in Sambrook et al., supra.

Alternatively, the polynucleotides of the invention can be produced by chemical synthesis. Such synthesis methods are described below.

It is contemplated that the nucleotide sequences presented herein may contain some small percentage of errors. These errors may arise in the normal course of determination of nucleotide sequences. Sequence errors can be corrected by obtaining seeds deposited under the accession numbers cited above, propagating them, isolating genomic DNA or appropriate mRNA from the resulting plants or seeds thereof, amplifying the relevant fragment of the genomic DNA or mRNA using primers having a sequence that flanks the erroneous sequence, and sequencing the amplification product.

I.A. Probes, Primers and Substrates

SDFs of the invention can be applied to substrates for use in array applications such as, but not limited to, assays of global gene expression, for example under varying conditions of development, growth conditions. The arrays can also be used in diagnostic or forensic methods (WO95/35505, US 5,445,943 and US 5,410,270).

5 Probes and primers of the instant invention will hybridize to a polynucleotide comprising a sequence in Tables 1 and 2. Though many different nucleotide sequences can encode an amino acid sequence, the sequences of Tables 1 and 2 are generally preferred for encoding polypeptides of the invention. However, the sequence of the probes and/or primers of the instant invention need not be identical to those in Tables 1 and 2 or the complements
10 thereof. For example, some variation in probe or primer sequence and/or length can allow additional family members to be detected, as well as orthologous genes and more taxonomically distant related sequences. Similarly, probes and/or primers of the invention can include additional nucleotides that serve as a label for detecting the formed duplex or for subsequent cloning purposes.

15 Probe length will vary depending on the application. For use as primers, probes are 12-40 nucleotides, preferably 18-30 nucleotides long. For use in mapping, probes are preferably 50 to 500 nucleotides, preferably 100-250 nucleotides long. For Southern hybridizations, probes as long as several kilobases can be used as explained below.

20 The probes and/or primers can be produced by synthetic procedures such as the triester method of Matteucci et al. *J. Am. Chem. Soc.* 103:3185(1981); or according to Urdea et al. *Proc. Natl. Acad.* 80:7461 (1981) or using commercially available automated oligonucleotide synthesizers.

25 I.B. Methods of Detection and Isolation

The polynucleotides of the invention can be utilized in a number of methods known to those skilled in the art as probes and/or primers to isolate and detect polynucleotides, including, without limitation: Southern, Northern, Branched DNA hybridization assays, polymerase chain reaction, and microarray assays, and variations thereof. Specific methods
30 given by way of examples, and discussed below include:

Hybridization

Methods of Mapping

Southern Blotting

Isolating cDNA from Related Organisms

Isolating and/or Identifying Orthologous Genes.

Also, the nucleic acid molecules of the invention can be used in other methods, such as high density oligonucleotide hybridizing assays, described, for example, in U.S. Pat. Nos.

5 6,004,753; 5,945,306; 5,945,287; 5,945,308; 5,919,686; 5,919,661; 5,919,627; 5,874,248; 5,871,973; 5,871,971; and 5,871,930; and PCT Pub. Nos. WO 9946380; WO 9933981; WO 9933870; WO 9931252; WO 9915658; WO 9906572; WO 9858052; WO 9958672; and WO 9810858.

10 B.1. Hybridization

The isolated SDFs of Tables 1 and 2 of the present invention can be used as probes and/or primers for detection and/or isolation of related polynucleotide sequences through hybridization. Hybridization of one nucleic acid to another constitutes a physical property that defines the subject SDF of the invention and the identified related sequences. Also, such hybridization imposes structural limitations on the pair. A good general discussion of the factors for determining hybridization conditions is provided by Sambrook et al. ("Molecular Cloning, a Laboratory Manual, 2nd ed., c. 1989 by Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY; *see esp.*, chapters 11 and 12). Additional considerations and details of the physical chemistry of hybridization are provided by G.H. Keller and M.M. Manak "DNA Probes", 2nd Ed. pp. 1-25, c. 1993 by Stockton Press, New York, NY.

20 Depending on the stringency of the conditions under which these probes and/or primers are used, polynucleotides exhibiting a wide range of similarity to those in Tables 1 and 2 can be detected or isolated. When the practitioner wishes to examine the result of membrane hybridizations under a variety of stringencies, an efficient way to do so is to perform the hybridization under a low stringency condition, then to wash the hybridization membrane under increasingly stringent conditions.

30 When using SDFs to identify orthologous genes in other species, the practitioner will preferably adjust the amount of target DNA of each species so that, as nearly as is practical, the same number of genome equivalents are present for each species examined. This prevents faint signals from species having large genomes, and thus small numbers of genome equivalents per mass of DNA, from erroneously being interpreted as absence of the corresponding gene in the genome.

The probes and/or primers of the instant invention can also be used to detect or isolate nucleotides that are "identical" to the probes or primers. Two nucleic acid sequences or polypeptides are said to be "identical" if the sequence of nucleotides or amino acid residues, respectively, in the two sequences is the same when aligned for maximum correspondence as described below.

Isolated polynucleotides within the scope of the invention also include allelic variants of the specific sequences presented in Tables 1 and 2. The probes and/or primers of the invention can also be used to detect and/or isolate polynucleotides exhibiting at least 80% sequence identity with the sequences of Tables 1 and 2 or fragments thereof.

With respect to nucleotide sequences, degeneracy of the genetic code provides the possibility to substitute at least one base of the base sequence of a gene with a different base without causing the amino acid sequence of the polypeptide produced from the gene to be changed. Hence, the DNA of the present invention may also have any base sequence that has been changed from a sequence in Tables 1 and 2 by substitution in accordance with degeneracy of genetic code. References describing codon usage include: Carels *et al.*, *J. Mol. Evol.* 46: 45 (1998) and Fennoy *et al.*, *Nucl. Acids Res.* 21(23): 5294 (1993).

B.2. Mapping

The isolated SDF DNA of the invention can be used to create various types of genetic and physical maps of the genome of corn, Arabidopsis, soybean, rice, wheat, or other plants. Some SDFs may be absolutely associated with particular phenotypic traits, allowing construction of gross genetic maps. While not all SDFs will immediately be associated with a phenotype, all SDFs can be used as probes for identifying polymorphisms associated with phenotypes of interest. Briefly, one method of mapping involves total DNA isolation from individuals. It is subsequently cleaved with one or more restriction enzymes, separated according to mass, transferred to a solid support, hybridized with SDF DNA and the pattern of fragments compared. Polymorphisms associated with a particular SDF are visualized as differences in the size of fragments produced between individual DNA samples after digestion with a particular restriction enzyme and hybridization with the SDF. After identification of polymorphic SDF sequences, linkage studies can be conducted. By using the individuals showing polymorphisms as parents in crossing programs, F2 progeny recombinants or recombinant inbreds, for example, are then analyzed. The order of DNA

polymorphisms along the chromosomes can be determined based on the frequency with which they are inherited together versus independently. The closer two polymorphisms are together in a chromosome the higher the probability that they are inherited together. Integration of the relative positions of all the polymorphisms and associated marker SDFs can produce a genetic map of the species, where the distances between markers reflect the recombination frequencies in that chromosome segment.

The use of recombinant inbred lines for such genetic mapping is described for *Arabidopsis* by Alonso-Blanco et al. (*Methods in Molecular Biology*, vol.82, "Arabidopsis Protocols", pp. 137-146, J.M. Martinez-Zapater and J. Salinas, eds., c. 1998 by Humana Press, Totowa, NJ) and for corn by Burr ("Mapping Genes with Recombinant Inbreds", pp. 249-254. In Freeling, M. and V. Walbot (Ed.), *The Maize Handbook*, c. 1994 by Springer-Verlag New York, Inc.: New York, NY, USA; Berlin Germany; Burr et al. *Genetics* (1998) 118: 519; Gardiner, J. et al., (1993) *Genetics* 134: 917). This procedure, however, is not limited to plants and can be used for other organisms (such as yeast) or for individual cells.

The SDFs of the present invention can also be used for simple sequence repeat (SSR) mapping. Rice SSR mapping is described by Morgante et al. (*The Plant Journal* (1993) 3: 165), Panaud et al. (*Genome* (1995) 38: 1170); Senior et al. (*Crop Science* (1996) 36: 1676), Taramino et al. (*Genome* (1996) 39: 277) and Ahn et al. (*Molecular and General Genetics* (1993) 241: 483-90). SSR mapping can be achieved using various methods. In one instance, polymorphisms are identified when sequence specific probes contained within an SDF flanking an SSR are made and used in polymerase chain reaction (PCR) assays with template DNA from two or more individuals of interest. Here, a change in the number of tandem repeats between the SSR-flanking sequences produces differently sized fragments (U.S. Patent 5,766,847). Alternatively, polymorphisms can be identified by using the PCR fragment produced from the SSR-flanking sequence specific primer reaction as a probe against Southern blots representing different individuals (U.H. Refseth et al., (1997) *Electrophoresis* 18: 1519).

Genetic and physical maps of crop species have many uses. For example, these maps can be used to devise positional cloning strategies for isolating novel genes from the mapped crop species. In addition, because the genomes of closely related species are largely syntenic (that is, they display the same ordering of genes within the genome), these maps can be used to isolate novel alleles from relatives of crop species by positional cloning strategies.

The various types of maps discussed above can be used with the SDFs of the invention to identify Quantitative Trait Loci (QTLs). Many important crop traits, such as the solids content of tomatoes, are quantitative traits and result from the combined interactions of several genes. These genes reside at different loci in the genome, oftentimes on different
5 chromosomes, and generally exhibit multiple alleles at each locus. The SDFs of the invention can be used to identify QTLs and isolate specific alleles as described by de Vicente and Tanksley (*Genetics* 134:585 (1993)). In addition to isolating QTL alleles in present crop species, the SDFs of the invention can also be used to isolate alleles from the corresponding QTL of wild relatives. Transgenic plants having various combinations of QTL alleles can
10 then be created and the effects of the combinations measured. Once a desired allele combination has been identified, crop improvement can be accomplished either through biotechnological means or by directed conventional breeding programs (for review see Tanksley and McCouch, *Science* 277:1063 (1997)).

In another embodiment, the SDFs can be used to help create physical maps of the genome of corn, *Arabidopsis* and related species. Where SDFs have been ordered on a
15 genetic map, as described above, they can be used as probes to discover which clones in large libraries of plant DNA fragments in YACs, BACs, etc. contain the same SDF or similar sequences, thereby facilitating the assignment of the large DNA fragments to chromosomal positions. Subsequently, the large BACs, YACs, etc. can be ordered unambiguously by more
20 detailed studies of their sequence composition (e.g. Marra et al. (1997) *Genomic Research* 7:1072-1084) and by using their end or other sequences to find the identical sequences in other cloned DNA fragments. The overlapping of DNA sequences in this way allows large contigs of plant sequences to be built that, when sufficiently extended, provide a complete physical map of a chromosome. Sometimes the SDFs themselves will provide the means of
25 joining cloned sequences into a contig.

The patent publication WO95/35505 and U.S. Patents 5,445,943 and 5,410,270 describe scanning multiple alleles of a plurality of loci using hybridization to arrays of oligonucleotides. These techniques are useful for each of the types of mapping discussed above.

30 Following the procedures described above and using a plurality of the SDFs of the present invention, any individual can be genotyped. These individual genotypes can be used for the identification of particular cultivars, varieties, lines, ecotypes and genetically

modified plants or can serve as tools for subsequent genetic studies involving multiple phenotypic traits.

B.3 Southern Blot Hybridization

5 The sequences from Tables 1 and 2 can be used as probes for various hybridization techniques. These techniques are useful for detecting target polynucleotides in a sample or for determining whether transgenic plants, seeds or host cells harbor a gene or sequence of interest and thus might be expected to exhibit a particular trait or phenotype.

10 In addition, the SDFs from the invention can be used to isolate additional members of gene families from the same or different species and/or orthologous genes from the same or different species. This is accomplished by hybridizing an SDF to, for example, a Southern blot containing the appropriate genomic DNA or cDNA. Given the resulting hybridization data, one of ordinary skill in the art could distinguish and isolate the correct DNA fragments by size, restriction sites, sequence and stated hybridization conditions from a gel or from a library.

15 Identification and isolation of orthologous genes from closely related species and alleles within a species is particularly desirable because of their potential for crop improvement. Many important crop traits, such as the solid content of tomatoes, result from the combined interactions of the products of several genes residing at different loci in the genome. Generally, alleles at each of these loci can make quantitative differences to the trait. By identifying and isolating numerous alleles for each locus from within or different species, transgenic plants with various combinations of alleles can be created and the effects of the combinations measured. Once a more favorable allele combination has been identified, crop improvement can be accomplished either through biotechnological means or by directed
20 conventional breeding programs (Tanksley et al. *Science* 277:1063(1997)).

25 The results from hybridizations of the SDFs of the invention to, for example, Southern blots containing DNA from another species can also be used to generate restriction fragment maps for the corresponding genomic regions. These maps provide additional information about the relative positions of restriction sites within fragments, further
30 distinguishing mapped DNA from the remainder of the genome.

Physical maps can be made by digesting genomic DNA with different combinations of restriction enzymes.

Probes for Southern blotting to distinguish individual restriction fragments can range in size from 15 to 20 nucleotides to several thousand nucleotides. More preferably, the probe is 100 to 1,000 nucleotides long for identifying members of a gene family when it is found that repetitive sequences would complicate the hybridization. For identifying an entire
5 corresponding gene in another species, the probe is more preferably the length of the gene, typically 2,000 to 10,000 nucleotides, but probes 50-1,000 nucleotides long might be used. Some genes, however, might require probes up to 1,500 nucleotides long or overlapping probes constituting the full-length sequence to span their lengths.

Also, while it is preferred that the probe be homogeneous with respect to its sequence,
10 it is not necessary. For example, as described below, a probe representing members of a gene family having diverse sequences can be generated using PCR to amplify genomic DNA or RNA templates using primers derived from SDFs that include sequences that define the gene family.

For identifying corresponding genes in another species, the next most preferable
15 probe is a cDNA spanning the entire coding sequence, which allows all of the mRNA-coding fragment of the gene to be identified. Probes for Southern blotting can easily be generated from SDFs by making primers having the sequence at the ends of the SDF and using corn or *Arabidopsis* genomic DNA as a template. In instances where the SDF includes sequence conserved among species, primers including the conserved sequence can be used for PCR
20 with genomic DNA from a species of interest to obtain a probe.

Similarly, if the SDF includes a domain of interest, that fragment of the SDF can be used to make primers and, with appropriate template DNA, used to make a probe to identify genes containing the domain. Alternatively, the PCR products can be resolved, for example by gel electrophoresis, and cloned and/or sequenced. Using Southern hybridization, the variants of
25 the domain among members of a gene family, both within and across species, can be examined.

B.4.1 Isolating DNA from Related Organisms

The SDFs of the invention can be used to isolate the corresponding DNA from other organisms. Either cDNA or genomic DNA can be isolated. For isolating genomic DNA, a
30 lambda, cosmid, BAC or YAC, or other large insert genomic library from the plant of interest can be constructed using standard molecular biology techniques as described in detail by Sambrook et al. 1989 (Molecular Cloning: A Laboratory Manual, 2nd ed. Cold Spring Harbor

Laboratory Press, New York) and by Ausubel et al. 1992 (Current Protocols in Molecular Biology, Greene Publishing, New York).

To screen a phage library, for example, recombinant lambda clones are plated out on appropriate bacterial medium using an appropriate *E. coli* host strain. The resulting plaques are lifted from the plates using nylon or nitrocellulose filters. The plaque lifts are processed through denaturation, neutralization, and washing treatments following the standard protocols outlined by Ausubel et al. (1992). The plaque lifts are hybridized to either radioactively labeled or non-radioactively labeled SDF DNA at room temperature for about 16 hours, usually in the presence of 50% formamide and 5X SSC (sodium chloride and sodium citrate) buffer and blocking reagents. The plaque lifts are then washed at 42°C with 1% Sodium Dodecyl Sulfate (SDS) and at a particular concentration of SSC. The SSC concentration used is dependent upon the stringency at which hybridization occurred in the initial Southern blot analysis performed. For example, if a fragment hybridized under medium stringency (e.g., $T_m - 20^\circ\text{C}$), then this condition is maintained or preferably adjusted to a less stringent condition (e.g., $T_m - 30^\circ\text{C}$) to wash the plaque lifts. Positive clones show detectable hybridization e.g., by exposure to X-ray films or chromogen formation. The positive clones are then subsequently isolated for purification using the same general protocol outlined above. Once the clone is purified, restriction analysis can be conducted to narrow the region corresponding to the gene of interest. The restriction analysis and succeeding subcloning steps can be done using procedures described by, for example Sambrook et al. (1989) cited above.

The procedures outlined for the lambda library are essentially similar to those used for YAC library screening, except that the YAC clones are harbored in bacterial colonies. The YAC clones are plated out at reasonable density on nitrocellulose or nylon filters supported by appropriate bacterial medium in petri plates. Following the growth of the bacterial clones, the filters are processed through the denaturation, neutralization, and washing steps following the procedures of Ausubel et al. 1992. The same hybridization procedures for lambda library screening are followed.

To isolate cDNA, similar procedures using appropriately modified vectors are employed. For instance, the library can be constructed in a lambda vector appropriate for cloning cDNA such as $\lambda\text{gt}11$. Alternatively, the cDNA library can be made in a plasmid vector. cDNA for cloning can be prepared by any of the methods known in the art, but is

preferably prepared as described above. Preferably, a cDNA library will include a high proportion of full-length clones.

B. 5. Isolating and/or Identifying Orthologous Genes

Probes and primers of the invention can be used to identify and/or isolate polynucleotides related to those in Tables 1 and 2. Related polynucleotides are those that are native to other plant organisms and exhibit either similar sequence or encode polypeptides with similar biological activity. One specific example is an orthologous gene. Orthologous genes have the same functional activity. As such, orthologous genes may be distinguished from homologous genes. The percentage of identity is a function of evolutionary separation and, in closely related species, the percentage of identity can be 98 to 100%. The amino acid sequence of a protein encoded by an orthologous gene can be less than 75% identical, but tends to be at least 75% or at least 80% identical, more preferably at least 90%, most preferably at least 95% identical to the amino acid sequence of the reference protein.

To find orthologous genes, the probes are hybridized to nucleic acids from a species of interest under low stringency conditions, preferably one where sequences containing as much as 40-45% mismatches will be able to hybridize. This condition is established by $T_m - 40^\circ\text{C}$ to $T_m - 48^\circ\text{C}$ (see below). Blots are then washed under conditions of increasing stringency. It is preferable that the wash stringency be such that sequences that are 85 to 100% identical will hybridize. More preferably, sequences 90 to 100% identical will hybridize and most preferably only sequences greater than 95% identical will hybridize. One of ordinary skill in the art will recognize that, due to degeneracy in the genetic code, amino acid sequences that are identical can be encoded by DNA sequences as little as 67% identical or less. Thus, it is preferable, for example, to make an overlapping series of shorter probes, on the order of 24 to 45 nucleotides, and individually hybridize them to the same arrayed library to avoid the problem of degeneracy introducing large numbers of mismatches.

As evolutionary divergence increases, genome sequences also tend to diverge. Thus, one of skill will recognize that searches for orthologous genes between more divergent species will require the use of lower stringency conditions compared to searches between closely related species. Also, degeneracy of the genetic code is more of a problem for searches in the genome of a species more distant evolutionarily from the species that is the source of the SDF probe sequences.

Therefore the method described in Bouckaert et al., U.S. Ser. No. 60/121,700 Atty. Dkt. No. 2750-117P, Client Dkt. No. 00010.001, filed February 25, 1999, hereby incorporated in its entirety by reference, can be applied to the SDFs of the present invention to isolate related genes from plant species which do not hybridize to the corn *Arabidopsis*, soybean, rice, wheat, and other plant sequences of Tables 1 and 2.

Identification of the relationship of nucleotide or amino acid sequences among plant species can be done by comparing the nucleotide or amino acid sequences of SDFs of the present application with nucleotide or amino acid sequences of other SDFs such as those present in applications listed in the table below:

Country	Attorney No.	Client No.	Filed	Application No.
United States	2750-0301P	80002.001	9/4/1998	60/099,672
United States	2750-0300P	80001.001	9/4/1998	60/099,671
United States	2750-0302P	80003.001	9/11/1998	60/099,933
United States	2750-0304P	80004.001	9/17/1998	60/100,864
United States	2750-0305P	80005.001	9/18/1998	60/101,042
United States	2750-0306P	80006.001	9/21/1998	60/101,255
United States	2750-0307P	80007.001	9/24/1998	60/101,682
United States	2750-0308P	80008.001	9/30/1998	60/102,533
United States	2750-0309P	80009.001	9/30/1998	60/102,460
United States	2750-0310P	80010.001	10/5/1998	60/103,116
United States	2750-0311P	80011.001	10/5/1998	60/103,141
United States	2750-0312P	80012.001	10/6/1998	60/103,215
United States	2750-0313P	80013.001	10/8/1998	60/103,554
United States	2750-0314P	80014.001	10/9/1998	60/103,574
United States	2750-0315P	80015.001	10/13/1998	60/103,907
United States	2750-0316P	80016.001	10/14/1998	60/104,268
United States	2750-0317P	80017.001	10/16/1998	60/104,680
United States	2750-0318P	80018.001	10/19/1998	60/104,828
United States	2750-0319P	80019.001	10/20/1998	60/105,008
United States	2750-0320P	80020.001	10/21/1998	60/105,142
United States	2750-0321P	80021.001	10/22/1998	60/105,533
United States	2750-0322P	80022.001	10/26/1998	60/105,571
United States	2750-0323P	80023.001	10/27/1998	60/105,815
United States	2750-0324P	80024.001	10/29/1998	60/106,105
United States	2750-0325P	80025.001	10/30/1998	60/106,218
United States	2750-0326P	80026.001	11/2/1998	60/106,685
United States	2750-0327P	80027.001	11/6/1998	60/107,282
United States	2750-0329P	80029.001	11/9/1998	60/107,719
United States	2750-0328P	80028.001	11/9/1998	60/107,720
United States	2750-0330P	80030.001	11/10/1998	60/107,836
United States	2750-0331P	80031.001	11/12/1998	60/108,190
United States	2750-0332P	80032.001	11/16/1998	60/108,526
United States	2750-0333P	80033.001	11/17/1998	60/108,901
United States	2750-0334P	80034.001	11/19/1998	60/109,124
United States	2750-0335P	80035.001	11/19/1998	60/109,127

Country	Attorney No.	Client No.	Filed	Application No.
United States	2750-0336P	80036.001	11/20/1998	60/109,267
United States	2750-0337P	80037.001	11/23/1998	60/109,594
United States	2750-0338P	80038.001	11/25/1998	60/110,053
United States	2750-0339P	80039.001	11/25/1998	60/110,050
United States	2750-0340P	80040.001	11/27/1998	60/110,158
United States	2750-0341P	80041.001	11/30/1998	60/110,263
United States	2750-0342P	80042.001	12/1/1998	60/110,495
United States	2750-0343P	80043.001	12/2/1998	60/110,626
United States	2750-0344P	80044.001	12/3/1998	60/110,701
United States	2750-0345P	80045.001	12/7/1998	60/111,339
United States	2750-0346P	80046.001	12/9/1998	60/111,589
United States	2750-0347P	80047.001	12/10/1998	60/111,782
United States	2750-0348P	80048.001	12/11/1998	60/111,812
United States	2750-0349P	80049.001	12/14/1998	60/112,096
United States	2750-0350P	80050.001	12/15/1998	60/112,224
United States	2750-0351P	80051.001	12/16/1998	60/112,624
United States	2750-0352P	80052.001	12/17/1998	60/112,862
United States	2750-0353P	80053.001	12/18/1998	60/112,912
United States	2750-0354P	80054.001	12/21/1998	60/113,248
United States	2750-0355P	80055.001	12/22/1998	60/113,522
United States	2750-0356P	80056.001	12/23/1998	60/113,826
United States	2750-0357P	80057.001	12/28/1998	60/113,998
United States	2750-0358P	80058.001	12/29/1998	60/114,384
United States	2750-0359P	80059.001	12/30/1998	60/114,455
United States	2750-0360P	80060.001	1/4/1999	60/114,740
United States	2750-0361P	80061.001	1/6/1999	60/114,866
United States	2750-0365P	80065.001	1/7/1999	60/115,155
United States	2750-0366P	80066.001	1/7/1999	60/115,156
United States	2750-0367P	80067.001	1/7/1999	60/115,154
United States	2750-0364P	80064.001	1/7/1999	60/115,151
United States	2750-0362P	80062.001	1/7/1999	60/115,153
United States	2750-0363P	80063.001	1/7/1999	60/115,152
United States	2750-0370P	80070.001	1/8/1999	60/115,293
United States	2750-0369P	80069.001	1/8/1999	60/115,365
United States	2750-0368P	80068.001	1/8/1999	60/115,364
United States	2750-0371P	80071.001	1/11/1999	60/115,339
United States	2750-0372P	80072.001	1/12/1999	60/115,518
United States	2750-0373P	80073.001	1/13/1999	60/115,847
United States	2750-0374P	80074.001	1/14/1999	60/115,905
United States	2750-0375P	80075.001	1/15/1999	60/116,383
United States	2750-0376P	80076.001	1/15/1999	60/116,384
United States	2750-0377P	80077.001	1/19/1999	60/116,329
United States	2750-0378P	80078.001	1/19/1999	60/116,340
United States	2750-0379P	80079.001	1/21/1999	60/116,674
United States	2750-0380P	80080.001	1/21/1999	60/116,672
United States	2750-0382P	80082.001	1/22/1999	60/116,962
United States	2750-0381P	80081.001	1/22/1999	60/116,960
United States	2750-0383P	80083.001	1/28/1999	60/117,756
United States	2750-0384P	80084.001	2/3/1999	60/118,672

Country	Attorney No.	Client No.	Filed	Application No.
United States	2750-0385P	80085.001	2/4/1999	60/118,808
United States	2750-0386P	80086.001	2/5/1999	60/118,778
United States	2750-0387P	80087.001	2/8/1999	60/119,029
United States	2750-0388P	80088.001	2/9/1999	60/119,332
United States	2750-0389P	80089.001	2/10/1999	60/119,462
United States	2750-0391P	80091.001	2/12/1999	60/119,922
United States	2750-0393P	80093.001	2/16/1999	60/120,198
United States	2750-0392P	80092.001	2/16/1999	60/120,196
United States	2750-0394P	80094.001	2/18/1999	60/120,583
United States	2750-0395P	80095.001	2/22/1999	60/121,072
United States	2750-0396P	80096.001	2/23/1999	60/121,334
United States	2750-0397P	80097.001	2/24/1999	60/121,470
United States	2750-0398P	80098.001	2/25/1999	60/121,704
United States	2750-0390P	80090.001	2/25/1999	60/121,825
United States	2750-0399P	80099.001	2/26/1999	60/122,107
United States	2750-0400P	80100.001	3/1/1999	60/122,266
United States	2750-0401P	80101.001	3/2/1999	60/122,568
United States	2750-0402P	80102.001	3/3/1999	60/122,611
United States	2750-0403P	80103.001	3/4/1999	60/121,775
United States	2750-0404P	80104.001	3/5/1999	60/123,534
United States	2750-0405P	80105.001	3/5/1999	60/123,180
United States	2750-0406P	80106.001	3/9/1999	60/123,680
United States	2750-0407P	80107.001	3/9/1999	60/123,548
United States	2750-0408P	80108.001	3/10/1999	60/123,715
United States	2750-0409P	80109.001	3/10/1999	60/123,726
United States	2750-0410P	80110.001	3/11/1999	60/124,263
United States	2750-0411P	80111.001	3/12/1999	60/123,941
United States	2750-0412P	80112.001	3/23/1999	60/125,788
United States	2750-0413P	80113.001	3/25/1999	60/126,264
United States	2750-0414P	80114.001	3/29/1999	60/126,785
United States	2750-0415P	80115.001	4/1/1999	60/127,462
United States	2750-0416P	91000.001	4/6/1999	60/128,234
United States	2750-0417P	91001.001	4/8/1999	60/128,714
United States	2750-0418P	80118.001	4/16/1999	60/129,845
United States	2750-0420P	80120.001	4/19/1999	60/130,077
United States	2750-0421P	80121.001	4/21/1999	60/130,449
United States	2750-0303P	80115.002	4/23/1999	60/130,510
United States	2750-0422P	80122.001	4/23/1999	60/130,891
United States	2750-0423P	80123.001	4/28/1999	60/131,449
United States	2750-0424P	80124.001	4/30/1999	60/132,407
United States	2750-0425P	80125.001	4/30/1999	60/132,048
United States	2750-0426P	80126.001	5/4/1999	60/132,484
United States	2750-0427P	80127.001	5/5/1999	60/132,485
United States	2750-0428P	91002.001	5/6/1999	60/132,487
United States	2750-0429P	80129.001	5/6/1999	60/132,486
United States	2750-0430P	80130.001	5/7/1999	60/132,863
United States	2750-0431P	80131.001	5/11/1999	60/134,256
United States	2750-0433P	00025.001	5/14/1999	60/134,221
United States	2750-0432P	91006.001	5/14/1999	60/134,370

Country	Attorney No.	Client No.	Filed	Application No.
United States	2750-0434P	80116.001	5/14/1999	60/134,219
United States	2750-0435P	80117.001	5/14/1999	60/134,218
United States	2750-0436P	91007.001	5/18/1999	60/134,768
United States	2750-0437P	91008.001	5/19/1999	60/134,941
United States	2750-0438P	91009.001	5/20/1999	60/135,124
United States	2750-0439P	91010.001	5/21/1999	60/135,353
United States	2750-0440P	91011.001	5/24/1999	60/135,629
United States	2750-0441P	91012.001	5/25/1999	60/136,021
United States	2750-0442P	91013.001	5/27/1999	60/136,392
United States	2750-0444P	91014.001	5/28/1999	60/136,782
United States	2750-0445P	91015.001	6/1/1999	60/137,222
United States	2750-0446P	91016.001	6/3/1999	60/137,528
United States	2750-0447P	91017.001	6/4/1999	60/137,502
United States	2750-0449P	91018.001	6/7/1999	60/137,724
United States	2750-0450P	91019.001	6/8/1999	60/138,094
United States	2750-0458P	00033.002	6/10/1999	60/138,847
United States	2750-0457P	00033.001	6/10/1999	60/138,540
United States	2750-0463P	00034.001	6/14/1999	60/139,119
United States	2750-0462P	80132.012	6/16/1999	60/139,452
United States	2750-0461P	80132.011	6/16/1999	60/139,453
United States	2750-0464P	00037.001	6/17/1999	60/139,492
United States	2750-0453P	80132.005	6/18/1999	60/139,462
United States	2750-0466P	00039.001	6/18/1999	60/139,750
United States	2750-0460P	80132.010	6/18/1999	60/139,455
United States	2750-0465P	00038.001	6/18/1999	60/139,763
United States	2750-0456P	80132.008	6/18/1999	60/139,456
United States	2750-0454P	80132.006	6/18/1999	60/139,457
United States	2750-0452P	80132.004	6/18/1999	60/139,461
United States	2750-0451P	80132.003	6/18/1999	60/139,459
United States	2750-0448P	80132.002	6/18/1999	60/139,454
United States	2750-0443P	80132.001	6/18/1999	60/139,458
United States	2750-0455P	80132.007	6/18/1999	60/139,460
United States	2750-0459P	80132.009	6/18/1999	60/139,463
United States	2750-0467P	00042.001	6/21/1999	60/139,817
United States	2750-0468P	00043.001	6/22/1999	60/139,899
United States	2750-0470P	00042.002	6/23/1999	60/140,353
United States	2750-0469P	00044.001	6/23/1999	60/140,354
United States	2750-0471P	00045.001	6/24/1999	60/140,695
United States	2750-0472P	00046.001	6/28/1999	60/140,823
United States	2750-0473P	00048.001	6/29/1999	60/140,991
United States	2750-0474P	00049.001	6/30/1999	60/141,287
United States	2750-0475P	00050.001	7/1/1999	60/141,842
United States	2750-0476P	00051.001	7/1/1999	60/142,154
United States	2750-0477P	00052.001	7/2/1999	60/142,055
United States	2750-0478P	00053.001	7/6/1999	60/142,390
United States	2750-0479P	00054.001	7/8/1999	60/142,803
United States	2750-0480P	00058.001	7/9/1999	60/142,920
United States	2750-0481P	00059.001	7/12/1999	60/142,977
United States	2750-0482P	00060.001	7/13/1999	60/143,542

Country	Attorney No.	Client No.	Filed	Application No.
United States	2750-0489P	00061.001	7/14/1999	60/143,624
United States	2750-0490P	00062.001	7/15/1999	60/144,005
United States	2750-0486P	80134.004	7/16/1999	60/144,085
United States	2750-0485P	80134.003	7/16/1999	60/144,086
United States	2750-0495P	80134.013	7/19/1999	60/144,335
United States	2750-0494P	80134.010	7/19/1999	60/144,333
United States	2750-0497P	00064.001	7/19/1999	60/144,325
United States	2750-0492P	80134.008	7/19/1999	60/144,331
United States	2750-0488P	80134.006	7/19/1999	60/144,332
United States	2750-0496P	80134.014	7/19/1999	60/144,334
United States	2750-0499P	80134.012	7/20/1999	60/144,352
United States	2750-0502P	80135.002	7/20/1999	60/144,884
United States	2750-0500P	00065.001	7/20/1999	60/144,632
United States	2750-0503P	00066.001	7/21/1999	60/144,814
United States	2750-0483P	80134.001	7/21/1999	60/145,088
United States	2750-0484P	80134.002	7/21/1999	60/145,086
United States	2750-0493P	80134.009	7/22/1999	60/145,087
United States	2750-0491P	80134.007	7/22/1999	60/145,085
United States	2750-0487P	80134.005	7/22/1999	60/145,089
United States	2750-0504P	00067.001	7/22/1999	60/145,192
United States	2750-0498P	80134.011	7/23/1999	60/145,145
United States	2750-0501P	80135.001	7/23/1999	60/145,224
United States	2750-0505P	00069.001	7/23/1999	60/145,218
United States	2750-0506P	00070.001	7/26/1999	60/145,276
United States	2750-0507P	80136.001	7/27/1999	60/145,918
United States	2750-0509P	00071.001	7/27/1999	60/145,913
United States	2750-0508P	80136.002	7/27/1999	60/145,919
United States	2750-0510P	00072.001	7/28/1999	60/145,951
United States	2750-0511P	80137.001	8/2/1999	60/146,388
United States	2750-0512P	80137.002	8/2/1999	60/146,389
United States	2750-0513P	00073.001	8/2/1999	60/146,386
United States	2750-0514P	00074.001	8/3/1999	60/147,038
United States	2750-0517P	80138.002	8/4/1999	60/147,302
United States	2750-0515P	00076.001	8/4/1999	60/147,204
United States	2750-0518P	00077.001	8/5/1999	60/147,260
United States	2750-0519P	80136.003	8/5/1999	60/147,192
United States	2750-0520P	00079.001	8/6/1999	60/147,416
United States	2750-0516P	80138.001	8/6/1999	60/147,303
United States	2750-0521P	00080.001	8/9/1999	60/147,493
United States	2750-0523P	80139.002	8/9/1999	60/147,935
United States	2750-0522P	80139.001	8/10/1999	60/148,171
United States	2750-0524P	00081.001	8/11/1999	60/148,319
United States	2750-0526P	80141.002	8/12/1999	60/148,342
United States	2750-0527P	80141.003	8/12/1999	60/148,340
United States	2750-0530P	00082.001	8/12/1999	60/148,341
United States	2750-0528P	80141.004	8/12/1999	60/148,337
United States	2750-0525P	80141.001	8/12/1999	60/148,347
United States	2750-0529P	00083.001	8/13/1999	60/148,565
United States	2750-0532P	80142.002	8/13/1999	60/148,684

Country	Attorney No.	Client No.	Filed	Application No.
United States	2750-0531P	80142.001	8/16/1999	60/149,368
United States	2750-0536P	80001.005	8/17/1999	60/149,925
United States	2750-0537P	00084.001	8/17/1999	60/149,175
United States	2750-0535P	80001.004	8/17/1999	60/149,926
United States	2750-0534P	80001.003	8/17/1999	60/149,928
United States	2750-0533P	80001.002	8/17/1999	60/149,927
United States	2750-0538P	00085.001	8/18/1999	60/149,426
United States	2750-0541P	80143.002	8/20/1999	60/149,929
United States	2750-0542P	00087.001	8/20/1999	60/149,723
United States	2750-0539P	00086.001	8/20/1999	60/149,722
United States	2750-0540P	80143.001	8/23/1999	60/149,930
United States	2750-0543P	00088.001	8/23/1999	60/149,902
United States	2750-0544P	00089.001	8/25/1999	60/150,566
United States	2750-0547P	00090.001	8/26/1999	60/150,884
United States	2750-0545P	80144.001	8/27/1999	60/151,065
United States	2750-0548P	00091.001	8/27/1999	60/151,080
United States	2750-0546P	80144.002	8/27/1999	60/151,066
United States	2750-0549P	00092.001	8/30/1999	60/151,303
United States	2750-0552P	00093.001	8/31/1999	60/151,438
United States	2750-0553P	00094.001	9/1/1999	60/151,930
United States	2750-0550P	80001.006	9/3/1999	09/391,631
International	2750-0551F(PC)	80001.100	9/3/1999	99/204,38
United States	2750-0554P	00095.001	9/7/1999	60/152,363
United States	2750-0555P	00096.001	9/10/1999	60/153,070
United States	2750-0556P	00098.001	9/13/1999	60/153,758
United States	2750-0557P	00099.001	9/15/1999	60/154,018
United States	2750-0558P	00101.001	9/16/1999	60/154,039
United States	2750-0559P	00102.001	9/20/1999	60/154,779
United States	2750-0560P	00103.001	9/22/1999	60/155,139
United States	2750-0561P	00104.001	9/23/1999	60/155,486
United States	2750-0562P	00105.001	9/24/1999	60/155,659
United States	2750-0563P	00106.001	9/28/1999	60/156,458
United States	2750-0564P	00107.001	9/29/1999	60/156,596
United States	2750-0570P	00108.001	10/4/1999	60/157,117
United States	2750-0571P	00109.001	10/5/1999	60/157,753
International	2750-0569F(PC)	80010.102	10/5/1999	99/228,53
International	2750-0568F(PC)	80010.101	10/5/1999	99/228,54
United States	2750-0565P	80010.002	10/5/1999	09/413,198
United States	2750-0566P	80010.003	10/5/1999	09/412,922
International	2750-0567F(PC)	80010.100	10/5/1999	99/228,55
United States	2750-0572P	00110.001	10/6/1999	60/157,865
United States	2750-0575P	00111.001	10/7/1999	60/158,029
United States	2750-0576P	00112.001	10/8/1999	60/158,232
United States	2750-0577P	00113.001	10/12/1999	60/158,369
United States	2750-0583P	80148.002	10/13/1999	60/159,294
United States	2750-0579P	80146.002	10/13/1999	60/159,293
United States	2750-0574P	80145.002	10/13/1999	60/159,295
United States	2750-0578P	80146.001	10/14/1999	60/159,331
United States	2750-0582P	80148.001	10/14/1999	60/159,329

CONFIDENTIAL

Country	Attorney No.	Client No.	Filed	Application No.
United States	2750-0581P	80147.002	10/14/1999	60/159,637
United States	2750-0580P	80147.001	10/14/1999	60/159,638
United States	2750-0573P	80145.001	10/14/1999	60/159,330
United States	2750-0584P	00116.001	10/18/1999	60/159,584
United States	2750-0586P	80149.001	10/21/1999	60/160,814
United States	2750-0590P	80150.002	10/21/1999	60/160,767
United States	2750-0589P	80150.001	10/21/1999	60/160,768
United States	2750-0587P	80149.002	10/21/1999	60/160,770
United States	2750-0585P	00118.001	10/21/1999	60/160,815
United States	2750-0588P	00119.001	10/21/1999	60/160,741
United States	2750-0591P	00120.001	10/22/1999	60/160,980
United States	2750-0593P	80151.002	10/22/1999	60/160,981
United States	2750-0592P	80151.001	10/22/1999	60/160,989
United States	2750-0594P	00121.001	10/25/1999	60/161,405
United States	2750-0595P	80152.001	10/25/1999	60/161,406
United States	2750-0596P	80152.002	10/25/1999	60/161,404
United States	2750-0598P	80153.001	10/26/1999	60/161,360
United States	2750-0597P	00122.001	10/26/1999	60/161,361
United States	2750-0599P	80153.002	10/26/1999	60/161,359
United States	2750-0602P	80154.001	10/28/1999	60/161,992
United States	2750-0603P	80154.002	10/28/1999	60/161,993
United States	2750-0601P	00123.001	10/28/1999	60/161,920
United States	2750-0600P	80026.002	10/28/1999	09/428,944
United States	2750-0604P	00124.001	10/29/1999	60/162,143
United States	2750-0605P	80155.001	10/29/1999	60/162,142
United States	2750-0606P	80155.002	10/29/1999	60/162,228
United States	2750-0609P	80156.002	11/1/1999	60/162,895
United States	2750-0608P	80156.001	11/1/1999	60/162,891
United States	2750-0607P	00125.001	11/1/1999	60/162,894
United States	2750-0612P	80157.002	11/2/1999	60/163,091
United States	2750-0610P	00126.001	11/2/1999	60/163,093
United States	2750-0611P	80157.001	11/2/1999	60/163,092
United States	2750-0613P	00127.001	11/3/1999	60/163,249
United States	2750-0615P	80158.002	11/3/1999	60/163,281
United States	2750-0614P	80158.001	11/3/1999	60/163,248
United States	2750-0618P	80159.002	11/4/1999	60/163,380
United States	2750-0617P	80159.001	11/4/1999	60/163,381
United States	2750-0616P	00128.001	11/4/1999	60/163,379
United States	2750-0619P	00129.001	11/8/1999	60/164,146
United States	2750-0620P	80160.001	11/8/1999	60/164,151
United States	2750-0621P	80160.002	11/8/1999	60/164,150
United States	2750-0623P	80161.002	11/9/1999	60/164,260
United States	2750-0625P	80162.002	11/9/1999	60/164,259
United States	2750-0628P	00131.001	11/10/1999	60/164,544
United States	2750-0627P	80163.002	11/10/1999	60/164,318
United States	2750-0626P	80163.001	11/10/1999	60/164,321
United States	2750-0629P	80164.001	11/10/1999	60/164,545
United States	2750-0622P	80161.001	11/10/1999	60/164,319
United States	2750-0630P	80164.002	11/10/1999	60/164,548

Country	Attorney No.	Client No.	Filed	Application No.
United States	2750-0624P	80162.001	11/10/1999	60/164,317
United States	2750-0631P	00132.001	11/12/1999	60/164,961
United States	2750-0632P	80165.001	11/12/1999	60/164,871
United States	2750-0633P	80165.002	11/12/1999	60/164,960
United States	2750-0634P	00133.001	11/12/1999	60/164,870
United States	2750-0635P	80166.001	11/12/1999	60/164,959
United States	2750-0636P	80166.002	11/12/1999	60/164,962
United States	2750-0639P	80167.002	11/15/1999	60/164,926
United States	2750-0638P	80167.001	11/15/1999	60/164,929
United States	2750-0637P	00134.001	11/15/1999	60/164,927
United States	2750-0642P	80168.002	11/16/1999	60/165,661
United States	2750-0641P	80168.001	11/16/1999	60/165,671
United States	2750-0640P	00135.001	11/16/1999	60/165,669
United States	2750-0643P	00136.001	11/17/1999	60/165,919
United States	2750-0644P	80169.001	11/17/1999	60/165,918
United States	2750-0645P	80169.002	11/17/1999	60/165,911
United States	2750-0648P	80170.002	11/18/1999	60/166,158
United States	2750-0646P	00137.001	11/18/1999	60/166,157
United States	2750-0647P	80170.001	11/18/1999	60/166,173
United States	2750-0651P	80171.002	11/19/1999	60/166,412
United States	2750-0649P	00139.001	11/19/1999	60/166,419
United States	2750-0650P	80171.001	11/19/1999	60/166,411
United States	2750-0652P	00140.001	11/22/1999	60/166,733
United States	2750-0653P	80172.001	11/22/1999	60/166,750
United States	2750-0655P	80173.002	11/23/1999	60/167,362
United States	2750-0658P	80174.002	11/24/1999	60/167,235
United States	2750-0654P	80173.001	11/24/1999	60/167,382
United States	2750-0656P	00141.001	11/24/1999	60/167,233
United States	2750-0657P	80174.001	11/24/1999	60/167,234
United States	2750-0659P	00142.001	11/30/1999	60/167,904
United States	2750-0661P	80175.002	11/30/1999	60/167,902
United States	2750-0660P	80175.001	11/30/1999	60/167,908
United States	2750-0664P	80176.001	12/1/1999	60/168,233
United States	2750-0662P	80042.002	12/1/1999	09/451,320
United States	2750-0665P	80176.002	12/1/1999	60/168,231
United States	2750-0663P	00143.001	12/1/1999	60/168,232
United States	2750-0667P	80177.001	12/2/1999	60/168,549
United States	2750-0666P	00144.001	12/2/1999	60/168,546
United States	2750-0668P	80177.002	12/2/1999	60/168,548
United States	2750-0670P	80178.001	12/3/1999	60/168,673
United States	2750-0669P	00145.001	12/3/1999	60/168,675
United States	2750-0671P	80178.002	12/3/1999	60/168,674
United States	2750-0673P	80179.001	12/7/1999	60/169,278
United States	2750-0674P	80179.002	12/7/1999	60/169,302
United States	2750-0672P	00147.001	12/7/1999	60/169,298
United States	2750-0676P	80180.002	12/8/1999	60/169,691
United States	2750-0675P	80180.001	12/8/1999	60/169,692
United States	2750-0677P	00149.001	12/16/1999	60/171,107
United States	2750-0679P	80181.002	12/16/1999	60/171,098

Country	Attorney No.	Client No.	Filed	Application No.
United States	2750-0678P	80181.001	12/16/1999	60/171,114
United States	2750-0683P	80060.002	1/4/2000	09/478,081
International	2750-0686F(PC)	80070.100	1/7/2000	00/004,66
United States	2750-0684P	80070.002	1/7/2000	09/479,221
United States	2750-0688P	80184.002	1/19/2000	60/176,910
United States	2750-0681P	80182.002	1/19/2000	60/176,866
United States	2750-0685P	80183.002	1/19/2000	60/176,867
United States	2750-0689P	00152.001	1/26/2000	60/178,166
United States	2750-0691P	80185.001	1/27/2000	60/177,666
United States	2750-0682P	80183.001	1/27/2000	60/178,546
United States	2750-0680P	80182.001	1/27/2000	60/178,544
United States	2750-0690P	00153.001	1/27/2000	60/178,547
United States	2750-0687P	80184.001	1/27/2000	60/178,545
United States	2750-0692P	00155.001	1/28/2000	60/178,754
United States	2750-0693P	80186.001	1/28/2000	60/178,755
United States	2750-0695P	00157.001	2/1/2000	60/179,395
United States	2750-0696P	80187.001	2/1/2000	60/179,388
United States	2750-0694P	80084.002	2/3/2000	09/497,191
United States	2750-0697P	00158.001	2/3/2000	60/180,039
United States	2750-0698P	80188.001	2/3/2000	60/180,139
United States	2750-0699P	00159.001	2/4/2000	60/180,206
United States	2750-0700P	80189.001	2/4/2000	60/180,207
United States	2750-0701P	00160.001	2/7/2000	60/180,695
United States	2750-0702P	80190.001	2/7/2000	60/180,696
United States	2750-0704P	80191.001	2/9/2000	60/181,214
United States	2750-0703P	00161.001	2/9/2000	60/181,228
United States	2750-0705P	00162.001	2/10/2000	60/181,476
United States	2750-0706P	80192.001	2/10/2000	60/181,551
United States	2750-0707P	00163.001	2/15/2000	60/182,477
United States	2750-0708P	80193.001	2/15/2000	60/182,516
United States	2750-0712P	00164.001	2/15/2000	60/182,512
United States	2750-0713P	80194.001	2/15/2000	60/182,478
United States	2750-0714P	00165.001	2/17/2000	60/183,166
United States	2750-0715P	80195.001	2/17/2000	60/183,165
United States	2750-0717P	80196.001	2/24/2000	60/184,658
United States	2750-0716P	00167.001	2/24/2000	60/184,667
United States	2750-0709P	80090.002	2/25/2000	09/513,996
United States	2750-0718P	91022.001	2/25/2000	60/185,140
United States	2750-0720P	80197.001	2/25/2000	60/185,119
Mexico	2750-0709F(MX)	80090.101	2/25/2000	00/001,973
Europe	2750-0709F(EP)	80090.103	2/25/2000	00/301,439
Canada	2750-0709F(CA)	80090.102	2/25/2000	23/006,92
United States	2750-0719P	00168.001	2/25/2000	60/185,118
United States	2750-0721P	91023.001	2/28/2000	60/185,398
United States	2750-0722P	00169.001	2/28/2000	60/185,396
United States	2750-0723P	80198.001	2/28/2000	60/185,397
United States	2750-0724P	91024.001	2/29/2000	60/185,750
United States	2750-0710P	80100.002	3/1/2000	09/517,537
United States	2750-0725P	00170.001	3/1/2000	

Country	Attorney No.	Client No.	Filed	Application No.
United States	2750-0726P	80199.001	3/1/2000	60/186,296
United States	2750-0727P	91025.001	3/1/2000	60/186,277
United States	2750-0728P	80200.001	3/2/2000	60/187,178
United States	2750-0729P	00172.001	3/2/2000	60/186,386
United States	2750-0711P	00171.001	3/2/2000	60/186,390
United States	2750-0730P	80201.001	3/2/2000	60/186,387
United States	2750-0731P	91026.001	3/3/2000	60/186,670
United States	2750-0732P	00173.001	3/3/2000	60/186,748
United States	2750-0733P	80202.001	3/3/2000	60/186,669
United States	2750-0734P	00174.001	3/7/2000	60/187,378
United States	2750-0735P	91027.001	3/7/2000	60/187,379
United States	2750-0736P	00175.001	3/8/2000	60/187,896
United States	2750-0737P	80203.001	3/8/2000	60/187,888
United States	2750-0738P	91028.001	3/9/2000	60/187,985
United States	2750-0740P	80204.001	3/10/2000	60/188,186
United States	2750-0739P	00177.001	3/10/2000	60/188,187
United States	2750-0741P	91030.001	3/10/2000	
United States	2750-0742P	00178.001	3/10/2000	60/188,185
United States	2750-0743P	80205.001	3/10/2000	60/188,175
United States	2750-0744P	91031.001	3/13/2000	60/188,687
United States	2750-0745P	00179.001	3/14/2000	60/189,080
United States	2750-0746P	80206.001	3/14/2000	60/189,052
United States	2750-0748P	00180.001	3/15/2000	60/189,461
United States	2750-0749P	80207.001	3/15/2000	60/189,462
United States	2750-0747P	91032.001	3/15/2000	60/189,460
United States	2750-0757P	91034.001	3/16/2000	60/189,965
United States	2750-0755P	00181.001	3/16/2000	60/189,953
United States	2750-0753P	80211.001	3/16/2000	60/190,121
United States	2750-0752P	80210.001	3/16/2000	60/189,948
United States	2750-0751P	80209.001	3/16/2000	60/189,947
United States	2750-0750P	80208.001	3/16/2000	60/190,120
United States	2750-0756P	80212.001	3/16/2000	60/189,959
United States	2750-0754P	91033.001	3/16/2000	60/189,958
United States	2750-0762P	80214.001	3/20/2000	60/190,089
United States	2750-0761P	00183.001	3/20/2000	60/190,545
United States	2750-0760P	91035.001	3/20/2000	60/190,060
United States	2750-0758P	00182.001	3/20/2000	60/190,069
United States	2750-0759P	80213.001	3/20/2000	60/190,070
United States	2750-0764P	80215.001	3/22/2000	60/191,097
United States	2750-0763P	00184.001	3/22/2000	60/191,084
United States	2750-0766P	00185.001	3/23/2000	60/191,543
United States	2750-0767P	80216.001	3/23/2000	60/191,545
United States	2750-0765P	91036.001	3/23/2000	60/191,549
United States	2750-0769P	00186.001	3/24/2000	60/191,823
United States	2750-0770P	80217.001	3/24/2000	60/191,825
United States	2750-0768P	91037.001	3/24/2000	60/191,826
United States	2750-0771P	91038.001	3/27/2000	60/192,420
United States	2750-0772P	00187.001	3/27/2000	60/192,421
United States	2750-0773P	80218.001	3/27/2000	60/192,308

OFFICE OF THE ATTORNEY GENERAL

Country	Attorney No.	Client No.	Filed	Application No.
United States	2750-0774P	91039.001	3/29/2000	60/192,855
United States	2750-0775P	00188.001	3/29/2000	60/192,940
United States	2750-0776P	80219.001	3/29/2000	60/192,941
United States	2750-0778P	00189.001	3/30/2000	60/193,244
United States	2750-0779P	80220.001	3/30/2000	60/193,245
United States	2750-0777P	91040.001	3/30/2000	60/193,243
United States	2750-0781P	00190.001	3/31/2000	60/193,453
United States	2750-0782P	80221.001	3/31/2000	60/193,455
United States	2750-0780P	91041.001	3/31/2000	60/193,469
United States	2750-0787P	80222.001	4/4/2000	
United States	2750-0786P	00191.001	4/4/2000	
United States	2750-0785P	91042.001	4/4/2000	
United States	2750-0789P	91043.001	4/5/2000	
United States	2750-0790P	00192.001	4/5/2000	
United States	2750-0791P	80223.001	4/5/2000	
United States	2750-0792P	91044.001	4/5/2000	
United States	2750-0783P	91000.002	4/6/2000	
Mexico	2750-0783F(MX)	91000.100	4/6/2000	00/003,391
United States	2750-0796P	80225.001	4/6/2000	
Europe	2750-0783F(EP)	91000.101	4/6/2000	00/302,919
Canada	2750-0783F(CA)	91000.102	4/6/2000	
United States	2750-0784P	91045.001	4/6/2000	
United States	2750-0795P	00194.001	4/6/2000	
United States	2750-0793P	00193.001	4/6/2000	
United States	2750-0794P	80224.001	4/6/2000	60/194,872
United States	2750-0797P	91046.001	4/7/2000	
United States	2750-0798P	00195.001	4/7/2000	60/195,283
United States	2750-0799P	80226.001	4/7/2000	60/195,257
United States	2750-0804P	80228.001	4/11/2000	60/196,089
United States	2750-0803P	00196.001	4/11/2000	60/196,169
United States	2750-0802P	91047.001	4/11/2000	60/196,168
United States	2750-0801P	80227.002	4/11/2000	60/196,211
United States	2750-0808P	00200.001	4/12/2000	60/196,485
United States	2750-0805P	91048.001	4/12/2000	60/196,483
United States	2750-0800P	80227.001	4/12/2000	60/196,212
United States	2750-0807P	80229.001	4/12/2000	
United States	2750-0809P	80230.001	4/12/2000	60/196,486
United States	2750-0806P	00197.001	4/12/2000	60/196,487
United States	2750-0811P	80231.002	4/13/2000	60/196,213
United States	2750-0814P	91049.001	4/14/2000	60/197,397
United States	2750-0810P	80231.001	4/14/2000	
United States	2750-0816P	80233.001	4/17/2000	60/197,678
United States	2750-0817P	91050.001	4/17/2000	60/198,268
United States	2750-0818P	00202.001	4/17/2000	60/198,133
United States	2750-0813P	80232.002	4/17/2000	60/197,871
United States	2750-0812P	80232.001	4/17/2000	60/197,870
United States	2750-0819P	80234.001	4/17/2000	60/197,671
United States	2750-0815P	00201.001	4/17/2000	60/197,687
United States	2750-0820P	91051.001	4/19/2000	60/198,400

Country	Attorney No.	Client No.	Filed	Application No.
United States	2750-0821P	00203.001	4/19/2000	60/198,386
United States	2750-0822P	80235.001	4/19/2000	60/198,373
United States	2750-0823P	91052.001	4/20/2000	60/198,629
United States	2750-0824P	00204.001	4/20/2000	60/198,619
United States	2750-0825P	80236.001	4/20/2000	60/198,623
United States	2750-0827P	00206.001	4/21/2000	60/198,767
United States	2750-0828P	80237.001	4/21/2000	60/198,763
United States	2750-0826P	91053.001	4/21/2000	
United States	2750-0831P	80238.001	4/24/2000	60/199,122
United States	2750-0830P	00207.001	4/24/2000	
United States	2750-0829P	91054.001	4/24/2000	
United States	2750-0833P	92002.001	4/26/2000	
United States	2750-0832P	92001.001	4/26/2000	60/200,034
United States	2750-0834P	00208.001	4/26/2000	
United States	2750-0835P	80239.001	4/26/2000	60/199,818
United States	2750-0836P	00210.001	4/27/2000	60/200,103
United States	2750-0837P	80240.001	4/27/2000	
United States	2750-0788P	80123.002	4/28/2000	
United States	2750-0846P	80243.002	4/28/2000	
United States	2750-0844P	80242.002	4/28/2000	
United States	2750-0845P	80243.001	5/1/2000	60/201,016
United States	2750-0847P	80244.001	5/1/2000	
United States	2750-0843P	80242.001	5/1/2000	
United States	2750-0848P	80244.002	5/1/2000	
United States	2750-0839P	80241.001	5/1/2000	
United States	2750-0840P	91055.001	5/1/2000	60/200,885
United States	2750-0841P	92001.002	5/1/2000	
United States	2750-0842P	92002.002	5/1/2000	60/201,018
United States	2750-0850P	80245.001	5/2/2000	60/201,305
United States	2750-0849P	91056.001	5/2/2000	60/201,279
United States	2750-0838P	00211.001	5/2/2000	60/201,275
United States	2750-0856P	91057.001	5/4/2000	
United States	2750-0858P	80246.001	5/4/2000	
United States	2750-0852P	80126.002	5/4/2000	
United States	2750-0857P	00212.001	5/4/2000	60/201,740
United States	2750-0860P	00213.001	5/5/2000	60/202,112
Europe	2750-0851F(EP)	91002.101	5/5/2000	
Canada	2750-0851F(CA)	91002.100	5/5/2000	
Mexico	2750-0851F(MX)	91002.102	5/5/2000	
United States	2750-0861P	80247.001	5/5/2000	60/202,180
United States	2750-0859P	91058.001	5/5/2000	
United States	2750-0851P	91002.002	5/5/2000	
United States	2750-0855P	80130.002	5/5/2000	
United States	2750-0853P	80127.002	5/5/2000	
United States	2750-0854P	80129.002	5/5/2000	
United States	2750-0862P	00214.001	5/9/2000	
United States	2750-0865P	00215.001	5/9/2000	
United States	2750-0866P	80249.001	5/9/2000	
United States	2750-0864P	91059.001	5/9/2000	

Country	Attorney No.	Client No.	Filed	Application No.
United States	2750-0863P	80248.001	5/9/2000	09/572,408
United States	2750-0877P	91060.001	5/10/2000	
United States	2750-0878P	00216.001	5/10/2000	
United States	2750-0879P	80252.001	5/10/2000	
United States	2750-0882P	80253.001	5/11/2000	
United States	2750-0881P	00217.001	5/11/2000	
United States	2750-0880P	91061.001	5/11/2000	
United States	2750-0871P	80131.002	5/11/2000	
United States	2750-0868P	80250.002	5/11/2000	
United States	2750-0867P	80250.001	5/11/2000	
United States	2750-0870P	80251.002	5/11/2000	
United States	2750-0869P	80251.001	5/11/2000	
Europe	2750-0875F(EP)	91006.101	5/12/2000	
Mexico	2750-0875F(MX)	91006.102	5/12/2000	
United States	2750-0873P	00025.002	5/12/2000	
United States	2750-0874P	80116.002	5/12/2000	
United States	2750-0885P	80254.001	5/12/2000	
United States	2750-0872P	80117.002	5/12/2000	
United States	2750-0884P	00219.001	5/12/2000	
Canada	2750-0875F(CA)	91006.100	5/12/2000	
United States	2750-0883P	91062.001	5/12/2000	
United States	2750-0875P	91006.002	5/12/2000	
United States	2750-0888P	80255.001	5/15/2000	
United States	2750-0886P	91063.001	5/15/2000	
United States	2750-0887P	00220.001	5/15/2000	
United States	2750-0891P	00221.001	5/16/2000	
United States	2750-0892P	80256.001	5/16/2000	
United States	2750-0890P	92002.003	5/17/2000	
United States	2750-0893P	00222.001	5/17/2000	
United States	2750-0889P	92001.003	5/17/2000	
United States	2750-0894P	80257.001	5/17/2000	
Canada	2750-0876F(CA)	91007.100	5/18/2000	
Europe	2750-0876F(EP)	91007.101	5/18/2000	
Mexico	2750-0876F(MX)	91007.102	5/18/2000	
United States	2750-0876P	91007.002	5/18/2000	
United States	2750-0895P	00223.001	5/18/2000	
United States	2750-0896P	80258.001	5/18/2000	
United States	2750-0897P	00224.001	5/19/2000	
United States	2750-0898P	80259.001	5/19/2000	
United States	2750-0901P	80260.001	5/22/2000	
United States	2750-0900P	00225.001	5/22/2000	
United States	2750-0899P	91064.001	5/22/2000	
United States	2750-0902P	00226.001	5/23/2000	
United States	2750-0903P	80261.001	5/23/2000	
United States	2750-0904P	00227.001	5/24/2000	
United States	2750-0905P	80262.001	5/24/2000	
United States	2750-0906P	91065.001	5/25/2000	
United States	2750-0911P	80264.001	5/26/2000	
United States	2750-0910P	00229.001	5/26/2000	

Country	Attorney No.	Client No.	Filed	Application No.
United States	2750-0907P	00228.001	5/26/2000	
United States	2750-0908P	80263.001	5/26/2000	
United States	2750-0909P	91066.001	5/26/2000	
United States	2750-0914P	80265.001	5/30/2000	
United States	2750-0913P	00230.001	5/30/2000	
United States	2750-0912P	91067.001	5/30/2000	
United States	2750-0921P	80268.001	6/1/2000	
United States	2750-0916P	80266.002	6/1/2000	
United States	2750-0918P	80267.002	6/1/2000	
United States	2750-0919P	91068.001	6/1/2000	
United States	2750-0920P	00231.001	6/1/2000	
United States	2750-0915P	80266.001	6/2/2000	
United States	2750-0917P	80267.001	6/2/2000	
United States	2750-0922P	91069.001	6/5/2000	
United States	2750-0923P	00232.001	6/5/2000	
United States	2750-0924P	80269.001	6/5/2000	
United States	2750-0925P	91070.001	6/5/2000	
United States	2750-0926P	00233.001	6/5/2000	
United States	2750-0927P	80270.001	6/5/2000	
United States	2750-0931P	80271.001	6/8/2000	
United States	2750-0930P	00234.001	6/8/2000	
United States	2750-0929P	91071.001	6/8/2000	
Canada	2750-0928F(CA)	00033.100	6/9/2000	
Mexico	2750-0928F(MX)	00033.102	6/9/2000	
United States	2750-0928P	00033.003	6/9/2000	09/592,459
United States	2750-0933P	80272.001	6/9/2000	
Mexico	2750-1037F(MX)		6/9/2000	
United States	2750-0932P	00235.001	6/9/2000	
Europe	2750-0928F(EP)	00033.101	6/12/2000	
United States	2750-0935P	00237.001	6/13/2000	
United States	2750-0936P	80273.001	6/13/2000	
United States	2750-0937P	91072.001	6/13/2000	
United States	2750-0934P	00034.002	6/14/2000	
United States	2750-0939P	80274.001	6/15/2000	
United States	2750-0940P	91074.001	6/15/2000	
United States	2750-0938P	00238.001	6/15/2000	
United States	2750-0952P	80132.021	6/16/2000	
United States	2750-0953P	80132.022	6/16/2000	
United States	2750-0955P	80132.024	6/16/2000	
United States	2750-0948P	80132.017	6/16/2000	
Europe	2750-0941F(EP)	00037.101	6/16/2000	
Mexico	2750-0941F(MX)	00037.102	6/16/2000	
United States	2750-0954P	80132.023	6/16/2000	
United States	2750-0947P	80132.016	6/16/2000	
United States	2750-0943P	00039.002	6/16/2000	
United States	2750-0944P	80132.013	6/16/2000	
United States	2750-0945P	80132.014	6/16/2000	
Canada	2750-0941F(CA)	00037.100	6/16/2000	
United States	2750-0946P	80132.015	6/16/2000	

Country	Attorney No.	Client No.	Filed	Application No.
United States	2750-0951P	80132.020	6/16/2000	
United States	2750-0942P	00038.002	6/16/2000	
United States	2750-0949P	80132.018	6/16/2000	
United States	2750-0950P	80132.019	6/16/2000	
United States	2750-0941P	00037.002	6/16/2000	
United States	2750-0956P	00239.001	6/19/2000	
United States	2750-0957P	80275.001	6/19/2000	
United States	2750-0958P	91075.001	6/19/2000	
United States	2750-0959P	00240.001	6/20/2000	
United States	2750-0961P	91076.001	6/20/2000	
United States	2750-0960P	80276.001	6/20/2000	
Canada	2750-0971F(CA)	00042.100	6/21/2000	
United States	2750-0971P	00042.003	6/21/2000	
Mexico	2750-0971F(MX)	00042.102	6/21/2000	
Europe	2750-0971F(EP)	00042.101	6/21/2000	
United States	2750-0966P	80278.001	6/22/2000	
United States	2750-0962P	00242.001	6/22/2000	
United States	2750-0963P	80277.001	6/22/2000	
United States	2750-0964P	91077.001	6/22/2000	
United States	2750-0967P	91079.001	6/22/2000	
United States	2750-0965P	00246.001	6/22/2000	
Canada	2750-0972F(CA)	00043.100	6/22/2000	
Europe	2750-0972F(EP)	00043.101	6/22/2000	
Mexico	2750-0972F(MX)	00043.102	6/22/2000	
United States	2750-0972P	00043.002	6/22/2000	
Canada	2750-0975F(CA)	00045.100	6/23/2000	
Europe	2750-0975F(EP)	00045.101	6/23/2000	
Mexico	2750-0973F(MX)	00044.102	6/23/2000	
United States	2750-0975P	00045.002	6/23/2000	
United States	2750-0973P	00044.002	6/23/2000	
Mexico	2750-0975F(MX)	00045.102	6/23/2000	
Europe	2750-0973F(EP)	00044.101	6/23/2000	
Canada	2750-0973F(CA)	00044.100	6/23/2000	
United States	2750-1036P	80280.001	6/27/2000	
United States	2750-1035P	00248.001	6/27/2000	
United States	2750-0970P	91080.001	6/27/2000	
United States	2750-0969P	80279.001	6/27/2000	
United States	2750-0968P	00247.001	6/27/2000	
Mexico	2750-0976F(MX)	00046.102	6/28/2000	
United States	2750-1039P	80281.001	6/28/2000	
United States	2750-0976P	00046.002	6/28/2000	
Canada	2750-0976F(CA)	00046.100	6/28/2000	
Europe	2750-0976F(EP)	00046.101	6/28/2000	
United States	2750-1038P	00249.001	6/28/2000	
Canada	2750-0977F(CA)	00048.100	6/29/2000	
Europe	2750-0977F(EP)	00048.101	6/29/2000	
United States	2750-0977P	00048.002	6/29/2000	
Mexico	2750-0977F(MX)	00048.102	6/29/2000	
Canada	2750-0980F(CA)	00051.100	6/30/2000	

Country	Attorney No.	Client No.	Filed	Application No.
Canada	2750-0981F(CA)	00052.100	6/30/2000	
Mexico	2750-0980F(MX)	00051.102	6/30/2000	
Europe	2750-0981F(EP)	00052.101	6/30/2000	
Europe	2750-0980F(EP)	00051.101	6/30/2000	
United States	2750-0981P	00052.002	6/30/2000	
United States	2750-0980P	00051.002	6/30/2000	
Mexico	2750-0981F(MX)	00052.102	6/30/2000	
United States	2750-0978P	00049.002	6/30/2000	
Canada	2750-0978F(CA)	00049.100	6/30/2000	
United States	2750-0979P	00050.002	6/30/2000	
Canada	2750-0979F(CA)	00050.100	6/30/2000	
Europe	2750-0979F(EP)	00050.101	6/30/2000	
Europe	2750-0978F(EP)	00049.101	6/30/2000	
United States	2750-1041P	80282.001	6/30/2000	
Mexico	2750-0978F(MX)	00049.102	6/30/2000	
Mexico	2750-0979F(MX)	00050.102	6/30/2000	
United States	2750-1040P	00250.001	6/30/2000	
United States	2750-1042P	00252.001	7/5/2000	
United States	2750-1043P	80283.001	7/5/2000	
United States	2750-0982P	00053.002	7/6/2000	
Mexico	2750-0982F(MX)	00053.102	7/6/2000	
Europe	2750-0982F(EP)	00053.101	7/6/2000	
Canada	2750-0982F(CA)	00053.100	7/6/2000	
United States	2750-0984P	00058.002	7/7/2000	
Mexico	2750-0984F(MX)	00058.102	7/7/2000	
Europe	2750-0984F(EP)	00058.101	7/7/2000	
Canada	2750-0984F(CA)	00058.100	7/7/2000	
United States	2750-0983P	00054.002	7/7/2000	
Canada	2750-0983F(CA)	00054.100	7/7/2000	
Europe	2750-0983F(EP)	00054.101	7/7/2000	
Mexico	2750-0983F(MX)	00054.102	7/7/2000	
United States	2750-1045P	00253.001	7/11/2000	
United States	2750-1046P	80284.001	7/11/2000	
United States	2750-1044P	91081.001	7/11/2000	
Canada	2750-0985F(CA)	00059.100	7/12/2000	
United States	2750-1052P	80287.002	7/12/2000	
United States	2750-1050P	80286.002	7/12/2000	
Mexico	2750-0985F(MX)	00059.102	7/12/2000	
United States	2750-0985P	00059.002	7/12/2000	
Europe	2750-0985F(EP)	00059.101	7/12/2000	
Europe	2750-0986F(EP)	00060.101	7/13/2000	
United States	2750-1054P	80288.002	7/13/2000	
Canada	2750-0986F(CA)	00060.100	7/13/2000	
Mexico	2750-0986F(MX)	00060.102	7/13/2000	
United States	2750-0986P	00060.002	7/13/2000	
Mexico	2750-0988F(MX)	00062.102	7/14/2000	
United States	2750-1061P	80134.018	7/14/2000	
United States	2750-1048P	80285.002	7/14/2000	
United States	2750-0987P	00061.002	7/14/2000	

Country	Attorney No.	Client No.	Filed	Application No.
Mexico	2750-0987F(MX)	00061.102	7/14/2000	
Europe	2750-0988F(EP)	00062.101	7/14/2000	
Canada	2750-0988F(CA)	00062.100	7/14/2000	
United States	2750-0988P	00062.002	7/14/2000	
Canada	2750-0987F(CA)	00061.100	7/14/2000	
Europe	2750-0987F(EP)	00061.101	7/14/2000	
United States	2750-1060P	80134.017	7/14/2000	
United States	2750-1055P	91082.001	7/18/2000	
United States	2750-1056P	00254.001	7/18/2000	
United States	2750-1057P	80291.001	7/18/2000	
Canada	2750-0989F(CA)	00064.100	7/19/2000	
United States	2750-1064P	80134.024	7/19/2000	
United States	2750-1062P	80134.020	7/19/2000	
United States	2750-1063P	80134.022	7/19/2000	
United States	2750-0989P	00064.002	7/19/2000	
Europe	2750-0989F(EP)	00064.101	7/19/2000	
Mexico	2750-0989F(MX)	00064.102	7/19/2000	
United States	2750-1066P	80135.004	7/20/2000	
United States	2750-0990P	00065.002	7/20/2000	
United States	2750-1065P	80134.026	7/20/2000	
Canada	2750-0990F(CA)	00065.100	7/20/2000	
Europe	2750-0990F(EP)	00065.101	7/20/2000	
Mexico	2750-0990F(MX)	00065.102	7/20/2000	
Europe	2750-0993F(EP)	00069.101	7/21/2000	
United States	2750-1073P	80134.025	7/21/2000	
United States	2750-1072P	80135.003	7/21/2000	
United States	2750-0992P	00067.002	7/21/2000	
Canada	2750-0992F(CA)	00067.100	7/21/2000	
Europe	2750-0992F(EP)	00067.101	7/21/2000	
United States	2750-1071P	80134.021	7/21/2000	
United States	2750-1067P	80134.015	7/21/2000	
Mexico	2750-0993F(MX)	00069.102	7/21/2000	
United States	2750-1070P	80134.019	7/21/2000	
United States	2750-1069P	80134.023	7/21/2000	
Mexico	2750-0991F(MX)	00066.102	7/21/2000	
Europe	2750-0991F(EP)	00066.101	7/21/2000	
United States	2750-1068P	80134.016	7/21/2000	
United States	2750-0991P	00066.002	7/21/2000	
Mexico	2750-0992F(MX)	00067.102	7/21/2000	
Canada	2750-0991F(CA)	00066.100	7/21/2000	
United States	2750-0993P	00069.002	7/21/2000	
Canada	2750-0993F(CA)	00069.100	7/21/2000	
United States	2750-1059P	00255.001	7/25/2000	
United States	2750-1081P	80293.001	7/25/2000	
United States	2750-1079P	80292.001	7/25/2000	
United States	2750-1058P	91083.001	7/25/2000	
United States	2750-1080P	00256.001	7/25/2000	
Mexico	2750-0994F(MX)	00070.102	7/26/2000	
United States	2750-0994P	00070.002	7/26/2000	

Country	Attorney No.	Client No.	Filed	Application No.
Canada	2750-0994F(CA)	00070.100	7/26/2000	
Europe	2750-0994F(EP)	00070.101	7/26/2000	
Mexico	2750-0995F(MX)	00071.102	7/27/2000	
United States	2750-0995P	00071.002	7/27/2000	
Canada	2750-0995F(CA)	00071.100	7/27/2000	
Europe	2750-0995F(EP)	00071.101	7/27/2000	
United States	2750-1074P	80136.004	7/27/2000	
United States	2750-1075P	80136.005	7/27/2000	
Canada	2750-0996F(CA)	00072.100	7/28/2000	
Europe	2750-0996F(EP)	00072.101	7/28/2000	
Mexico	2750-0996F(MX)	00072.102	7/28/2000	
United States	2750-0996P	00072.002	7/28/2000	

All applications listed in the table above are expressly incorporated herein by reference.

The SDFs of the invention can also be used as probes to search for genes that are related to the SDF within a species. Such related genes are typically considered to be members of a gene family. In such a case, the sequence similarity will often be concentrated into one or a few fragments of the sequence. The fragments of similar sequence that define the gene family typically encode a fragment of a protein or RNA that has an enzymatic or structural function. The percentage of identity in the amino acid sequence of the domain that defines the gene family is preferably at least 70%, more preferably 80 to 95%, most preferably 85 to 99%. To search for members of a gene family within a species, a low stringency hybridization is usually performed, but this will depend upon the size, distribution and degree of sequence divergence of domains that define the gene family. SDFs encompassing regulatory regions can be used to identify coordinately expressed genes by using the regulatory region sequence of the SDF as a probe.

In the instances where the SDFs are identified as being expressed from genes that confer a particular phenotype, then the SDFs can also be used as probes to assay plants of different species for those phenotypes.

I.C. Methods to Inhibit Gene Expression

The nucleic acid molecules of the present invention can be used to inhibit gene transcription and/or translation. Example of such methods include, without limitation:

Antisense Constructs;

Ribozyme Constructs;

Chimeraplast Constructs;
Co-Suppression;
Transcriptional Silencing; and
Other Methods of Gene Expression.

5

C.1 Antisense

In some instances it is desirable to suppress expression of an endogenous or exogenous gene. A well-known instance is the FLAVOR-SAVOR™ tomato, in which the gene encoding ACC synthase is inactivated by an antisense approach, thus delaying softening of the fruit after ripening. See for example, U.S. Patent No. 5,859,330; U.S. Patent No. 5,723,766; Oeller, et al, *Science*, 254:437-439(1991); and Hamilton et al, *Nature*, 346:284-287 (1990). Also, timing of flowering can be controlled by suppression of the *FLOWERING LOCUS C (FLC)*; high levels of this transcript are associated with late flowering, while absence of *FLC* is associated with early flowering (S.D. Michaels et al., *Plant Cell* 11:949 (1999). Also, the transition of apical meristem from production of leaves with associated shoots to flowering is regulated by *TERMINAL FLOWER1*, *APETALA1* and *LEAFY*. Thus, when it is desired to induce a transition from shoot production to flowering, it is desirable to suppress *TFL1* expression (S.J. Liljegren, *Plant Cell* 11:1007 (1999)). As another instance, arrested ovule development and female sterility result from suppression of the ethylene forming enzyme but can be reversed by application of ethylene (D. De Martinis et al., *Plant Cell* 11:1061 (1999)). The ability to manipulate female fertility of plants is useful in increasing fruit production and creating hybrids.

In the case of polynucleotides used to inhibit expression of an endogenous gene, the introduced sequence need not be perfectly identical to a sequence of the target endogenous gene. The introduced polynucleotide sequence will typically be at least substantially identical to the target endogenous sequence.

Some polynucleotide SDFs in Tables 1 and 2 represent sequences that are expressed in corn, wheat, rice, soybean *Arabidopsis* and/or other plants. Thus the invention includes using these sequences to generate antisense constructs to inhibit translation and/or degradation of transcripts of said SDFs, typically in a plant cell.

To accomplish this, a polynucleotide segment from the desired gene that can hybridize to the mRNA expressed from the desired gene (the “antisense segment”) is operably linked to a

promoter such that the antisense strand of RNA will be transcribed when the construct is present in a host cell. A regulated promoter can be used in the construct to control transcription of the antisense segment so that transcription occurs only under desired circumstances.

The antisense segment to be introduced generally will be substantially identical to at least a fragment of the endogenous gene or genes to be repressed. The sequence, however, need not be perfectly identical to inhibit expression. Further, the antisense product may hybridize to the untranslated region instead of or in addition to the coding sequence of the gene. The vectors of the present invention can be designed such that the inhibitory effect applies to other proteins within a family of genes exhibiting homology or substantial homology to the target gene.

For antisense suppression, the introduced antisense segment sequence also need not be full length relative to either the primary transcription product or the fully processed mRNA. Generally, a higher percentage of sequence identity can be used to compensate for the use of a shorter sequence. Furthermore, the introduced sequence need not have the same intron or exon pattern, and homology of non-coding segments may be equally effective. Normally, a sequence of between about 30 or 40 nucleotides and the full length of the transcript can be used, though a sequence of at least about 100 nucleotides is preferred, a sequence of at least about 200 nucleotides is more preferred, and a sequence of at least about 500 nucleotides is especially preferred.

C.2. Ribozymes

It is also contemplated that gene constructs representing ribozymes and based on the SDFs in Tables 1 and 2 are an object of the invention. Ribozymes can also be used to inhibit expression of genes by suppressing the translation of the mRNA into a polypeptide. It is possible to design ribozymes that specifically pair with virtually any target RNA and cleave the phosphodiester backbone at a specific location, thereby functionally inactivating the target RNA. In carrying out this cleavage, the ribozyme is not itself altered, and is thus capable of recycling and cleaving other molecules, making it a true enzyme. The inclusion of ribozyme sequences within antisense RNAs confers RNA-cleaving activity upon them, thereby increasing the activity of the constructs.

A number of classes of ribozymes have been identified. One class of ribozymes is derived from a number of small circular RNAs, which are capable of self-cleavage and replication in plants. The RNAs replicate either alone (viroid RNAs) or with a helper virus

(satellite RNAs). Examples include RNAs from avocado sunblotch viroid and the satellite RNAs from tobacco ringspot virus, lucerne transient streak virus, velvet tobacco mottle virus, solanum nodiflorum mottle virus and subterranean clover mottle virus. The design and use of target RNA-specific ribozymes is described in Haseloff et al. *Nature*, 334:585 (1988).

5 Like the antisense constructs above, the ribozyme sequence fragment necessary for pairing need not be identical to the target nucleotides to be cleaved, nor identical to the sequences in Tables 1 and 2. Ribozymes may be constructed by combining the ribozyme sequence and some fragment of the target gene which would allow recognition of the target gene mRNA by the resulting ribozyme molecule. Generally, the sequence in the ribozyme
10 capable of binding to the target sequence exhibits a percentage of sequence identity with at least 80%, preferably with at least 85%, more preferably with at least 90% and most preferably with at least 95%, even more preferably, with at least 96%, 97%, 98% or 99% sequence identity to some fragment of a sequence in Tables 1 and 2 or the complement thereof. The ribozyme can be equally effective in inhibiting mRNA translation by cleaving either in the untranslated or coding regions. Generally, a higher percentage of sequence identity can be used to
15 compensate for the use of a shorter sequence. Furthermore, the introduced sequence need not have the same intron or exon pattern, and homology of non-coding segments may be equally effective.

C.3. Chimeraplasts

20 The SDFs of the invention, such as those described by Tables 1 and 2, can also be used to construct chimeraplasts that can be introduced into a cell to produce at least one specific nucleotide change in a sequence corresponding to the SDF of the invention. A chimeraplast is an oligonucleotide comprising DNA and/or RNA that specifically hybridizes to a target region in a manner which creates a mismatched base-pair. This mismatched base-
25 pair signals the cell's repair enzyme machinery which acts on the mismatched region resulting in the replacement, insertion or deletion of designated nucleotide(s). The altered sequence is then expressed by the cell's normal cellular mechanisms. Chimeraplasts can be designed to repair mutant genes, modify genes, introduce site-specific mutations, and/or act to interrupt or alter normal gene function (US Pat. Nos. 6,010,907 and 6,004,804; and PCT
30 Pub. No. WO99/58723 and WO99/07865).

C.4. Sense Suppression

The SDFs of Tables 1 and 2 of the present invention are also useful to modulate gene expression by sense suppression. Sense suppression represents another method of gene suppression by introducing at least one exogenous copy or fragment of the endogenous sequence to be suppressed.

Introduction of expression cassettes in which a nucleic acid is configured in the sense orientation with respect to the promoter into the chromosome of a plant or by a self-replicating virus has been shown to be an effective means by which to induce degradation of mRNAs of target genes. For an example of the use of this method to modulate expression of endogenous genes see, Napoli et al., *The Plant Cell* 2:279 (1990), and U.S. Patents Nos. 5,034,323, 5,231,020, and 5,283,184. Inhibition of expression may require some transcription of the introduced sequence.

For sense suppression, the introduced sequence generally will be substantially identical to the endogenous sequence intended to be inactivated. The minimal percentage of sequence identity will typically be greater than about 65%, but a higher percentage of sequence identity might exert a more effective reduction in the level of normal gene products. Sequence identity of more than about 80% is preferred, though about 95% to absolute identity would be most preferred. As with antisense regulation, the effect would likely apply to any other proteins within a similar family of genes exhibiting homology or substantial homology to the suppressing sequence.

C.5. Transcriptional Silencing

The nucleic acid sequences of the invention, including the SDFs of Tables 1 and 2, and fragments thereof, contain sequences that can be inserted into the genome of an organism resulting in transcriptional silencing. Such regulatory sequences need not be operatively linked to coding sequences to modulate transcription of a gene. Specifically, a promoter sequence without any other element of a gene can be introduced into a genome to transcriptionally silence an endogenous gene (see, for example, Vaucheret, H et al. (1998) *The Plant Journal* 16: 651-659). As another example, triple helices can be formed using oligonucleotides based on sequences from Tables 1 and 2, fragments thereof, and substantially similar sequence thereto. The oligonucleotide can be delivered to the host cell and can bind to the promoter in the genome to form a triple helix and prevent transcription. An oligonucleotide of interest is one that can bind to the promoter and block binding of a transcription factor to the promoter. In such a case,

the oligonucleotide can be complementary to the sequences of the promoter that interact with transcription binding factors.

C.6. Other Methods to Inhibit Gene Expression

Yet another means of suppressing gene expression is to insert a polynucleotide into the gene of interest to disrupt transcription or translation of the gene.

Low frequency homologous recombination can be used to target a polynucleotide insert to a gene by flanking the polynucleotide insert with sequences that are substantially similar to the gene to be disrupted. Sequences from Tables 1 and 2, fragments thereof, and substantially similar sequence thereto can be used for homologous recombination.

In addition, random insertion of polynucleotides into a host cell genome can also be used to disrupt the gene of interest. Azpiroz-Leehan et al., *Trends in Genetics* **13**:152 (1997). In this method, screening for clones from a library containing random insertions is preferred to identifying those that have polynucleotides inserted into the gene of interest. Such screening can be performed using probes and/or primers described above based on sequences from Tables 1 and 2, fragments thereof, and substantially similar sequence thereto. The screening can also be performed by selecting clones or R₁ plants having a desired phenotype.

I.D. Methods of Functional Analysis

The constructs described in the methods under I.C. above can be used to determine the function of the polypeptide encoded by the gene that is targeted by the constructs.

Down-regulating the transcription and translation of the targeted gene in the host cell or organisms, such as a plant, may produce phenotypic changes as compared to a wild-type cell or organism. In addition, *in vitro* assays can be used to determine if any biological activity, such as calcium flux, DNA transcription, nucleotide incorporation, etc., are being modulated by the down-regulation of the targeted gene.

Coordinated regulation of sets of genes, e.g., those contributing to a desired polygenic trait, is sometimes necessary to obtain a desired phenotype. SDFs of the invention representing transcription activation and DNA binding domains can be assembled into hybrid transcriptional activators. These hybrid transcriptional activators can be used with their corresponding DNA elements (i.e., those bound by the DNA-binding SDFs) to effect

coordinated expression of desired genes (J.J. Schwarz et al., *Mol. Cell. Biol.* 12:266 (1992), A. Martinez et al., *Mol. Gen. Genet.* 261:546 (1999)).

The SDFs of the invention can also be used in the two-hybrid genetic systems to identify networks of protein-protein interactions (L. McAlister-Henn et al., *Methods* 19:330 (1999), J.C. Hu et al., *Methods* 20:80 (2000), M. Golovkin et al., *J. Biol. Chem.* 274:36428 (1999), K. Ichimura et al., *Biochem. Biophys. Res. Comm.* 253:532 (1998)). The SDFs of the invention can also be used in various expression display methods to identify important protein-DNA interactions (e.g. B. Luo et al., *J. Mol. Biol.* 266:479 (1997)).

I.E. Promoters

The SDFs of the invention are also useful as structural or regulatory sequences in a construct for modulating the expression of the corresponding gene in a plant or other organism, e.g. a symbiotic bacterium. For example, promoter sequences associated to SDFs of Tables 1 and 2 of the present invention can be useful in directing expression of coding sequences either as constitutive promoters or to direct expression in particular cell types, tissues, or organs or in response to environmental stimuli.

With respect to the SDFs of the present invention a promoter is likely to be a relatively small portion of a genomic DNA (gDNA) sequence located in the first 2000 nucleotides upstream from an initial exon identified in a gDNA sequence or initial "ATG" or methionine codon or translational start site in a corresponding cDNA sequence. Such promoters are more likely to be found in the first 1000 nucleotides upstream of an initial ATG or methionine codon or translational start site of a cDNA sequence corresponding to a gDNA sequence. In particular, the promoter is usually located upstream of the transcription start site. The fragments of a particular gDNA sequence that function as elements of a promoter in a plant cell will preferably be found to hybridize to gDNA sequences presented and described in Tables 1 and 2 at medium or high stringency, relevant to the length of the probe and its base composition.

Promoters are generally modular in nature. Promoters can consist of a basal promoter that functions as a site for assembly of a transcription complex comprising an RNA polymerase, for example RNA polymerase II. A typical transcription complex will include additional factors such as TF_{II}B, TF_{II}D, and TF_{II}E. Of these, TF_{II}D appears to be the only one to bind DNA directly. The promoter might also contain one or more enhancers and/or suppressors that

function as binding sites for additional transcription factors that have the function of modulating the level of transcription with respect to tissue specificity and of transcriptional responses to particular environmental or nutritional factors, and the like.

Short DNA sequences representing binding sites for proteins can be separated from each other by intervening sequences of varying length. For example, within a particular functional module, protein binding sites may be constituted by regions of 5 to 60, preferably 10 to 30, more preferably 10 to 20 nucleotides. Within such binding sites, there are typically 2 to 6 nucleotides that specifically contact amino acids of the nucleic acid binding protein. The protein binding sites are usually separated from each other by 10 to several hundred nucleotides, typically by 15 to 150 nucleotides, often by 20 to 50 nucleotides. DNA binding sites in promoter elements often display dyad symmetry in their sequence. Often elements binding several different proteins, and/or a plurality of sites that bind the same protein, will be combined in a region of 50 to 1,000 basepairs.

Elements that have transcription regulatory function can be isolated from their corresponding endogenous gene, or the desired sequence can be synthesized, and recombined in constructs to direct expression of a coding region of a gene in a desired tissue-specific, temporal-specific or other desired manner of inducibility or suppression. When hybridizations are performed to identify or isolate elements of a promoter by hybridization to the long sequences presented in Tables 1 and 2, conditions are adjusted to account for the above-described nature of promoters. For example short probes, constituting the element sought, are preferably used under low temperature and/or high salt conditions. When long probes, which might include several promoter elements are used, low to medium stringency conditions are preferred when hybridizing to promoters across species.

If a nucleotide sequence of an SDF, or part of the SDF, functions as a promoter or fragment of a promoter, then nucleotide substitutions, insertions or deletions that do not substantially affect the binding of relevant DNA binding proteins would be considered equivalent to the exemplified nucleotide sequence. It is envisioned that there are instances where it is desirable to decrease the binding of relevant DNA binding proteins to silence or down-regulate a promoter, or conversely to increase the binding of relevant DNA binding proteins to enhance or up-regulate a promoter and vice versa. In such instances, polynucleotides representing changes to the nucleotide sequence of the DNA-protein contact region by insertion of additional nucleotides, changes to identity of relevant nucleotides, including use of chemically-modified bases, or deletion of one or more nucleotides are

considered encompassed by the present invention. In addition, fragments of the promoter sequences described by Tables 1 and 2 and variants thereof can be fused with other promoters or fragments to facilitate transcription and/or transcription in specific type of cells or under specific conditions.

5 Promoter function can be assayed by methods known in the art, preferably by measuring activity of a reporter gene operatively linked to the sequence being tested for promoter function. Examples of reporter genes include those encoding luciferase, green fluorescent protein, GUS, neo, cat and bar.

I.F. UTRs and Junctions

10 Polynucleotides comprising untranslated (UTR) sequences and intron/exon junctions are also within the scope of the invention. UTR sequences include introns and 5' or 3' untranslated regions (5' UTRs or 3' UTRs). Fragments of the sequences shown in Tables 1 and 2 can comprise UTRs and intron/exon junctions.

These fragments of SDFs, especially UTRs, can have regulatory functions related to, for example, translation rate and mRNA stability. Thus, these fragments of SDFs can be isolated for use as elements of gene constructs for regulated production of polynucleotides encoding desired polypeptides.

15 Introns of genomic DNA segments might also have regulatory functions. Sometimes regulatory elements, especially transcription enhancer or suppressor elements, are found within introns. Also, elements related to stability of heteronuclear RNA and efficiency of splicing and of transport to the cytoplasm for translation can be found in intron elements. Thus, these segments can also find use as elements of expression vectors intended for use to transform plants.

20 Just as with promoters UTR sequences and intron/exon junctions can vary from those shown in Tables 1 and 2. Such changes from those sequences preferably will not affect the regulatory activity of the UTRs or intron/exon junction sequences on expression, transcription, or translation unless selected to do so. However, in some instances, down- or up-regulation of such activity may be desired to modulate traits or phenotypic or *in vitro* activity.

30 I.G. Coding Sequences

Isolated polynucleotides of the invention can include coding sequences that encode polypeptides comprising an amino acid sequence encoded by sequences in Tables 1 and 2 or an amino acid sequence presented in Tables 1 and 2.

A nucleotide sequence encodes a polypeptide if a cell (or a cell free *in vitro* system) expressing that nucleotide sequence produces a polypeptide having the recited amino acid sequence when the nucleotide sequence is transcribed and the primary transcript is subsequently processed and translated by a host cell (or a cell free *in vitro* system) harboring the nucleic acid. Thus, an isolated nucleic acid that encodes a particular amino acid sequence can be a genomic sequence comprising exons and introns or a cDNA sequence that represents the product of splicing thereof. An isolated nucleic acid encoding an amino acid sequence also encompasses heteronuclear RNA, which contains sequences that are spliced out during expression, and mRNA, which lacks those sequences.

Coding sequences can be constructed using chemical synthesis techniques or by isolating coding sequences or by modifying such synthesized or isolated coding sequences as described above.

In addition to coding sequences encoding the polypeptide sequences of Tables 1 and 2, which are native to corn, *Arabidopsis*, soybean, rice, wheat, and other plants the isolated polynucleotides can be polynucleotides that encode variants, fragments, and fusions of those native proteins. Such polypeptides are described below in part II.

In variant polynucleotides generally, the number of substitutions, deletions or insertions is preferably less than 20%, more preferably less than 15%; even more preferably less than 10%, 5%, 3% or 1% of the number of nucleotides comprising a particularly exemplified sequence. It is generally expected that non-degenerate nucleotide sequence changes that result in 1 to 10, more preferably 1 to 5 and most preferably 1 to 3 amino acid insertions, deletions or substitutions will not greatly affect the function of an encoded polypeptide. The most preferred embodiments are those wherein 1 to 20, preferably 1 to 10, most preferably 1 to 5 nucleotides are added to, deleted from and/or substituted in the sequences specifically disclosed in Tables 1 and 2.

Insertions or deletions in polynucleotides intended to be used for encoding a polypeptide preferably preserve the reading frame. This consideration is not so important in instances when the polynucleotide is intended to be used as a hybridization probe.

II. Polypeptides and Proteins

IIA. Native polypeptides and proteins

Polypeptides within the scope of the invention include both native proteins as well as
5 variants, fragments, and fusions thereof. Polypeptides of the invention are those encoded by
any of the six reading frames of sequences shown in Tables 1 and 2, preferably encoded by
the three frames reading in the 5' to 3' direction of the sequences as shown.

Native polypeptides include the proteins encoded by the sequences shown in Tables 1
and 2. Such native polypeptides include those encoded by allelic variants.

10 Polypeptide and protein variants will exhibit at least 75% sequence identity to those
native polypeptides of Tables 1 and 2. More preferably, the polypeptide variants will exhibit at
least 85% sequence identity; even more preferably, at least 90% sequence identity; more
preferably at least 95%, 96%, 97%, 98%, or 99% sequence identity. Fragments of polypeptide
or fragments of polypeptides will exhibit similar percentages of sequence identity to the
15 relevant fragments of the native polypeptide. Fusions will exhibit a similar percentage of
sequence identity in that fragment of the fusion represented by the variant of the native peptide.

Furthermore, polypeptide variants will exhibit at least one of the functional properties of
the native protein. Such properties include, without limitation, protein interaction, DNA
interaction, biological activity, immunological activity, receptor binding, signal transduction,
transcription activity, growth factor activity, secondary structure, three-dimensional structure,
20 etc. As to properties related to *in vitro* or *in vivo* activities, the variants preferably exhibit at least
60% of the activity of the native protein; more preferably at least 70%, even more preferably at
least 80%, 85%, 90% or 95% of at least one activity of the native protein.

One type of variant of native polypeptides comprises amino acid substitutions, deletions
25 and/or insertions. Conservative substitutions are preferred to maintain the function or activity of
the polypeptide.

Within the scope of percentage of sequence identity described above, a polypeptide of
the invention may have additional individual amino acids or amino acid sequences inserted into
the polypeptide in the middle thereof and/or at the N-terminal and/or C-terminal ends thereof.

30 Likewise, some of the amino acids or amino acid sequences may be deleted from the
polypeptide.

A.1 Antibodies

Isolated polypeptides can be utilized to produce antibodies. Polypeptides of the invention can generally be used, for example, as antigens for raising antibodies by known techniques. The resulting antibodies are useful as reagents for determining the distribution of the antigen protein within the tissues of a plant or within a cell of a plant. The antibodies are also useful for examining the production level of proteins in various tissues, for example in a wild-type plant or following genetic manipulation of a plant, by methods such as Western blotting.

Antibodies of the present invention, both polyclonal and monoclonal, may be prepared by conventional methods. In general, the polypeptides of the invention are first used to immunize a suitable animal, such as a mouse, rat, rabbit, or goat. Rabbits and goats are preferred for the preparation of polyclonal sera due to the volume of serum obtainable, and the availability of labeled anti-rabbit and anti-goat antibodies as detection reagents. Immunization is generally performed by mixing or emulsifying the protein in saline, preferably in an adjuvant such as Freund's complete adjuvant, and injecting the mixture or emulsion parenterally (generally subcutaneously or intramuscularly). A dose of 50-200 µg/injection is typically sufficient. Immunization is generally boosted 2-6 weeks later with one or more injections of the protein in saline, preferably using Freund's incomplete adjuvant. One may alternatively generate antibodies by *in vitro* immunization using methods known in the art, which for the purposes of this invention is considered equivalent to *in vivo* immunization.

Polyclonal antisera is obtained by bleeding the immunized animal into a glass or plastic container, incubating the blood at 25°C for one hour, followed by incubating the blood at 4°C for 2-18 hours. The serum is recovered by centrifugation (e.g., 1,000xg for 10 minutes). About 20-50 ml per bleed may be obtained from rabbits.

Monoclonal antibodies are prepared using the method of Kohler and Milstein, *Nature* 256: 495 (1975), or modification thereof. Typically, a mouse or rat is immunized as described above. However, rather than bleeding the animal to extract serum, the spleen (and optionally several large lymph nodes) is removed and dissociated into single cells. If desired, the spleen cells can be screened (after removal of nonspecifically adherent cells) by applying a cell suspension to a plate, or well, coated with the protein antigen. B-cells producing membrane-bound immunoglobulin specific for the antigen bind to the plate, and are not rinsed away with the rest of the suspension. Resulting B-cells, or all dissociated spleen cells, are then induced to fuse with myeloma cells to form hybridomas, and are cultured in a selective medium (e.g., hypoxanthine, aminopterin, thymidine medium, "HAT"). The resulting hybridomas are plated

by limiting dilution, and are assayed for the production of antibodies which bind specifically to the immunizing antigen (and which do not bind to unrelated antigens). The selected Mab-secreting hybridomas are then cultured either *in vitro* (e.g., in tissue culture bottles or hollow fiber reactors), or *in vivo* (as ascites in mice).

5 Other methods for sustaining antibody-producing B-cell clones, such as by EBV transformation, are known.

If desired, the antibodies (whether polyclonal or monoclonal) may be labeled using conventional techniques. Suitable labels include fluorophores, chromophores, radioactive atoms (particularly ^{32}P and ^{125}I), electron-dense reagents, enzymes, and ligands having specific binding
10 partners. Enzymes are typically detected by their activity. For example, horseradish peroxidase is usually detected by its ability to convert 3,3',5,5'-tetramethylbenzidine (TNB) to a blue pigment, quantifiable with a spectrophotometer.

A.2 In Vitro Applications of Polypeptides

Some polypeptides of the invention will have enzymatic activities that are useful *in vitro*.
5 For example, the soybean trypsin inhibitor (Kunitz) family is one of the numerous families of proteinase inhibitors. It comprises plant proteins which have inhibitory activity against serine proteinases from the trypsin and subtilisin families, thiol proteinases and aspartic proteinases. Thus, these peptides find *in vitro* use in protein purification protocols and perhaps in therapeutic settings requiring topical application of protease inhibitors.

20 Delta-aminolevulinic acid dehydratase (EC 4.2.1.24) (ALAD) catalyzes the second step in the biosynthesis of heme, the condensation of two molecules of 5-aminolevulinate to form porphobilinogen and is also involved in chlorophyll biosynthesis (Kaczor et al. (1994) Plant Physiol. 1-4: 1411-7; Smith (1988) Biochem. J. 249: 423-8; Schneider (1976) Z. naturforsch. [C] 31: 55-63). Thus, ALAD proteins can be used as catalysts in synthesis of
25 heme derivatives. Enzymes of biosynthetic pathways generally can be used as catalysts for *in vitro* synthesis of the compounds representing products of the pathway.

Polypeptides encoded by SDFs of the invention can be engineered to provide purification reagents to identify and purify additional polypeptides that bind to them. This allows one to identify proteins that function as multimers or elucidate signal transduction or
30 metabolic pathways. In the case of DNA binding proteins, the polypeptide can be used in a similar manner to identify the DNA determinants of specific binding (S. Pierrou et al., *Anal.*

Biochem. 229:99 (1995), S. Chusacultanachai et al., *J. Biol. Chem.* 274:23591 (1999), Q. Lin et al., *J. Biol. Chem.* 272:27274 (1997)).

II.B. POLYPEPTIDE VARIANTS, FRAGMENTS, AND FUSIONS

Generally, variants, fragments, or fusions of the polypeptides encoded by the maximum length sequence (MLS) can exhibit at least one of the activities of the identified domains and/or related polypeptides described in Sections (C) and (D) of Table 1 corresponding to the MLS of interest.

II.B.(1) Variants

A type of variant of the native polypeptides comprises amino acid substitutions.

Conservative substitutions, described above (see II.), are preferred to maintain the function or activity of the polypeptide. Such substitutions include conservation of charge, polarity, hydrophobicity, size, etc. For example, one or more amino acid residues within the sequence can be substituted with another amino acid of similar polarity that acts as a functional equivalent, for example providing a hydrogen bond in an enzymatic catalysis. Substitutes for an amino acid within an exemplified sequence are preferably made among the members of the class to which the amino acid belongs. For example, the nonpolar (hydrophobic) amino acids include alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan and methionine. The polar neutral amino acids include glycine, serine, threonine, cysteine, tyrosine, asparagine, and glutamine. The positively charged (basic) amino acids include arginine, lysine and histidine. The negatively charged (acidic) amino acids include aspartic acid and glutamic acid.

Within the scope of percentage of sequence identity described above, a polypeptide of the invention may have additional individual amino acids or amino acid sequences inserted into the polypeptide in the middle thereof and/or at the N-terminal and/or C-terminal ends thereof. Likewise, some of the amino acids or amino acid sequences may be deleted from the polypeptide. Amino acid substitutions may also be made in the sequences; conservative substitutions being preferred.

One preferred class of variants are those that comprise (1) the domain of an encoded polypeptide and/or (2) residues conserved between the encoded polypeptide and related polypeptides. For this class of variants, the encoded polypeptide sequence is changed by insertion, deletion, or substitution at positions flanking the domain and/or conserved residues.

Another class of variants includes those that comprise an encoded polypeptide sequence that is changed in the domain or conserved residues by a conservative substitution.

Yet another class of variants includes those that lack one of the *in vitro* activities, or structural features of the encoded polypeptides. One example is polypeptides or proteins produced from genes comprising dominant negative mutations. Such a variant may comprise an encoded polypeptide sequence with non-conservative changes in a particular domain or group of conserved residues.

II.A.(2) FRAGMENTS

Fragments of particular interest are those that comprise a domain identified for a polypeptide encoded by an MLS of the instant invention and variants thereof. Also, fragments that comprise at least one region of residues conserved between an MLS encoded polypeptide and its related polypeptides are of great interest. Fragments are sometimes useful as polypeptides corresponding to genes comprising dominant negative mutations are.

II.A.(3) FUSIONS

Of interest are chimeras comprising (1) a fragment of the MLS encoded polypeptide or variants thereof of interest and (2) a fragment of a polypeptide comprising the same domain. For example, an AP2 helix encoded by a MLS of the invention fused to second AP2 helix from ANT protein, which comprises two AP2 helices. The present invention also encompasses fusions of MLS encoded polypeptides, variants, or fragments thereof fused with related proteins or fragments thereof.

DEFINITION OF DOMAINS

The polypeptides of the invention may possess identifying domains as shown in Table 1. Specific domains within the MLS encoded polypeptides are indicated in Table 1. In addition, the domains within the MLS encoded polypeptide can be defined by the region that exhibits at least 70% sequence identity with the consensus sequences listed in the detailed description below of each of the domains.

The majority of the protein domain descriptions given below are obtained from Prosite, (<http://www.expasy.ch/prosite/>), and Pfam, (<http://pfam.wustl.edu/browse.shtml>).

1. (AAA) AAA-protein family signature

A large family of ATPases has been described [1 to 5] whose key feature is that they share a conserved region of about 220 amino acids that contains an ATP-binding site.

This family is now called AAA, for 'A'TPases 'A'ssociated with diverse cellular

'A'ctivities. The proteins that belong to this family either contain one or two AAA domains. Proteins containing two AAA domains:

- Mammalian and drosophila NSF (N-ethylmaleimide-sensitive fusion protein) and the fungal homolog, SEC18. These proteins are involved in intracellular transport between the endoplasmic reticulum and Golgi, as well as between different Golgi cisternae.
- Mammalian transitional endoplasmic reticulum ATPase (previously known as p97 or VCP) which is involved in the transfer of membranes from the endoplasmic reticulum to the golgi apparatus. This protein forms a ring-shaped homooligomer composed of six subunits. The yeast homolog is CDC48 and it may play a role in spindle pole proliferation.
- Yeast protein PAS1, essential for peroxisome assembly and the related protein PAS1 from *Pichia pastoris*.
- Yeast protein AFG2.
- *Sulfolobus acidocaldarius* protein SAV and *Halobacterium salinarium* cdcH which may be part of a transduction pathway connecting light to cell division.

Proteins containing a single AAA domain:

- *Escherichia coli* and other bacteria ftsH (or hflB) protein. FtsH is an ATP-dependent zinc metalloprotease that seems to degrade the heat-shock sigma-32 factor.

It is an integral membrane protein with a large cytoplasmic C-terminal domain that contain both the AAA and the protease domains.

- Yeast protein YME1, a protein important for maintaining the integrity of the mitochondrial compartment. YME1 is also a zinc-dependent protease.
- Yeast protein AFG3 (or YTA10). This protein also seems to contain a AAA domain followed by a zinc-dependent protease domain.

Subunits from the regulatory complex of the 26S proteasome [6] which is involved in the ATP-dependent degradation of ubiquitinated proteins:

- a) Mammalian subunit 4 and homologs in other higher eukaryotes, in yeast (gene YTA5) and fission yeast (gene mts2).

- b) Mammalian subunit 6 (TBP7) and homologs in other higher eukaryotes and in yeast (gene YTA2).
- c) Mammalian subunit 7 (MSS1) and homologs in other higher eukaryotes and in yeast (gene CIM5 or YTA3).
- 5 d) Mammalian subunit 8 (P45) and homologs in other higher eukaryotes and in yeast (SUG1 or CIM3 or TBY1) and fission yeast (gene let1).

Other probable subunits such as human TBP1 which seems to influences HIV gene expression by interacting with the virus tat transactivator protein and yeast YTA1 and YTA6.

- Yeast protein BCS1, a mitochondrial protein essential for the expression of the
- 10 Rieske iron-sulfur protein.
- Yeast protein MSP1, a protein involved in intramitochondrial sorting of proteins.
- Yeast protein PAS8, and the corresponding proteins PAS5 from *Pichia pastoris* and PAY4 from *Yarrowia lipolytica*.
- Mouse protein SKD1 and its fission yeast homolog (SpAC2G11.06).
- *Caenorhabditis elegans* meiotic spindle formation protein mei-1.
- Yeast protein SAP1.
- Yeast protein YTA7.
- *Mycobacterium leprae* hypothetical protein A2126A.

It is proposed that, in general, the AAA domains in these proteins act as ATP-
20 dependent protein clamps [5]. In addition to the ATP-binding 'A' and 'B' motifs, which are located in the N-terminal half of this domain, there is a highly conserved region located in the central part of the domain which was used to develop a signature pattern.

Consensus pattern: [LIVMT]-x-[LIVMT]-[LIVMF]-x-[GATMC]-[ST]-[NS]-x(4)-[LIVM]-
25 D-x-A-[LIFA]-x-R

[1] Froehlich K.-U., Fries H.W., Ruediger M., Erdmann R., Botstein D., Mecke D. J. Cell Biol. 114:443-453(1991).

[2] Erdmann R., Wiebel F.F., Flessau A., Rytka J., Beyer A., Froehlich K.-U., Kunau W.-H.
30 Cell 64:499-510(1991).

[3] Peters J.-M., Walsh M.J., Franke W.W. EMBO J. 9:1757-1767(1990).

[4] Kunau W.-H., Beyer A., Goette K., Marzioch M., Saidowsky J., Skaletz-Rorowski A., Wiebel F.F. Biochimie 75:209-224(1993).

[5] Confalonieri F., Duguet M. BioEssays 17:639-650(1995).[6] Hilt W., Wolf D.H. Trends Biochem. Sci. 21:96-102(1996).

2. ABC Membrane (ABC transporter transmembrane region). This family represents a unit of six transmembrane helices. Many members of the ABC transporter family (ABC tran) have two such regions. See also descriptions of ABC Tran, below, and ABC2 membrane, above.

3. (ABC Tran) ABC transporters family signature. On the basis of sequence similarities a family of related ATP-binding proteins has been characterized [1 to 5]. These proteins are associated with a variety of distinct biological processes in both prokaryotes and eukaryotes, but a majority of them are involved in active transport of small hydrophilic molecules across the cytoplasmic membrane. All these proteins share a conserved domain of some two hundred amino acid residues, which includes an ATP-binding site. These proteins are collectively known as ABC transporters. Proteins known to belong to this family are listed below (references are only provided for recently determined sequences). In prokaryotes: - Active transport systems components: alkylphosphonate uptake (phnC/phnK/ phnL); arabinose (araG); arginine (artP); dipeptide (dcjAD;dppD/dppF); ferric enterobactin (fepC); ferrichrome (fhuC); galactoside (mglA); glutamine (glnQ); glycerol-3-phosphate (ugpC); glycine betaine/L-proline (proV); glutamate/aspartate (gltL); histidine (hisP); iron(III) (sfuC), iron(III) dicitrate (fecE); lactose (lacK); leucine/isoleucine/valine (braF/braG; livF/livG); maltose (malK); molybdenum (modC); nickel (nikD/ nikE); oligopeptide (amiE/amiF; oppD/oppF); peptide (sapD/sapF); phosphate (pstB); putrescine (potG); ribose (rbsA); spermidine/putrescine (potA); sulfate (cysA); vitamin B12 (btuD). - Hemolysin/leukotoxin export proteins hlyB, cyaB and lktB. - Colicin V export protein cvaB. - Lactococcal export protein lcnC [6]. - Lantibiotic transport proteins nisT (nisin) and spaT (subtilin). - Extracellular proteases B and C export protein prtD. - Alkaline protease secretion protein aprD. - Beta-(1,2)-glucan export proteins chvA and ndvA. - Haemophilus influenzae capsule-polysaccharide export protein bexA. - Cytochrome c biogenesis proteins ccmA (also known as cycV and helA). - Polysialic acid transport protein kpsT. - Cell division associated ftsE protein (function unknown). - Copper processing protein nosF from Pseudomonas stutzeri. - Nodulation protein nodI from Rhizobium (function unknown). - Escherichia coli proteins cydC and cydD. - Subunit A of the ABC excision nuclease (gene uvrA). -

Erythromycin resistance protein from *Staphylococcus epidermidis* (gene *msrA*). - Tylosin resistance protein from *Streptomyces fradiae* (gene *tlrC*) [7]. - Heterocyst differentiation protein (gene *hetA*) from *Anabaena* PCC 7120. - Protein P29 from *Mycoplasma hyorhinis*, a probable component of a high affinity transport system. - *yhbG*, a putative protein whose gene is linked with *ntrA* in many bacteria such as *Escherichia coli*, *Klebsiella pneumoniae*, *Pseudomonas putida*, *Rhizobium meliloti* and *Thiobacillus ferrooxidans*. - *Escherichia coli* and related bacteria hypothetical proteins *yabJ*, *yadG*, *yagC*, *ybbA*, *ycjW*, *yddA*, *yehX*, *yejF*, *yheS*, *yhiG*, *yhiH*, *yjcW*, *yjjK*, *yojI*, *yrbF* and *ytfR*. In eukaryotes: - The multidrug transporters (Mdr) (P-glycoprotein), a family of closely related proteins which extrude a wide variety of drugs out of the cell (for a review see [8]). - Cystic fibrosis transmembrane conductance regulator (CFTR), which is most probably involved in the transport of chloride ions. - Antigen peptide transporters 1 (TAP1, PSF1, RING4, HAM-1, *mtp1*) and 2 (TAP2, PSF2, RING11, HAM-2, *mtp2*), which are involved in the transport of antigens from the cytoplasm to a membrane-bound compartment for association with MHC class I molecules. - 70 Kd peroxisomal membrane protein (PMP70). - ALDP, a peroxisomal protein involved in X-linked adrenoleukodystrophy [9]. - Sulfonylurea receptor [10], a putative subunit of the B-cell ATP-sensitive potassium channel. - *Drosophila* proteins white (*w*) and brown (*bw*), which are involved in the import of ommatidium screening pigments. - Fungal elongation factor 3 (EF-3). - Yeast STE6 which is responsible for the export of the α -factor pheromone. - Yeast mitochondrial transporter ATM1. - Yeast MDL1 and MDL2. - Yeast SNQ2. - Yeast sporidesmin resistance protein (gene *PDR5* or *STS1* or *YDR1*). - Fission yeast heavy metal tolerance protein *hmt1*. This protein is probably involved in the transport of metal-bound phytochelatins. - Fission yeast brefeldin A resistance protein (gene *bfr1* or *hba2*). - Fission yeast leptomycin B resistance protein (gene *pmd1*). - *mbpX*, a hypothetical chloroplast protein from Liverwort. - Prestalk-specific protein *tagB* from slime mold. This protein consists of two domains: a N-terminal subtilase catalytic domain and a C-terminal ABC transporter domain. As a signature pattern for this class of proteins, a conserved region which is located between the 'A' and the 'B' motifs of the ATP-binding site was used.

Consensus pattern: [LIVMFYC]-[SA]-[SAPGLVFYKQH]-G-[DENQMW]-[KRQASPCLIMFW]-[KRNQSTAVM]-[KRACLVM]-[LIVMFYPAN]-{PHY}-[LIVMFW]-[SAGCLIVP]-{FYWHP}-{KRHP}-[LIVMFYWSTA] The ATP-binding region is duplicated in *araG*, *mdl*, *msrA*, *rbsA*, *tlrC*, *uvrA*, *yejF*, Mdr's, CFTR, *pmd1* and in EF-3. In

some of those proteins, the above pattern only detect one of the two copies of the domain. The proteins belonging to this family also contain one or two copies of the ATP-binding motifs 'A' and 'B'.

- 5 [1] Higgins C.F., Hyde S.C., Mimmack M.M., Gileadi U., Gill D.R., Gallagher M.P. J. Bioenerg. Biomembr. 22:571-592(1990).
- [2] Higgins C.F., Gallagher M.P., Mimmack M.M., Pearce S.R. BioEssays 8:111-116(1988).
- [3] Higgins C.F., Hiles I.D., Salmond G.P.C., Gill D.R., Downie J.A., Evans I.J., Holland I.B., Gray L., Buckels S.D., Bell A.W., Hermodson M.A. Nature 323:448-450(1986).
- 10 [4] Doolittle R.F., Johnson M.S., Husain I., van Houten B., Thomas D.C., Sancar A. Nature 323:451-453(1986).
- [5] Blight M.A., Holland I.B. Mol. Microbiol. 4:873-880(1990).
- [6] Stoddard G.W., Petzel J.P., van Belkum M.J., Kok J., McKay L.L. Appl. Environ. Microbiol. 58:1952-1961(1992).
- 15 [7] Rosteck P.R. Jr., Reynolds P.A., Hershberger C.L. Gene 102:27-32(1991).
- [8] Gottesman M.M., Pastan I. J. Biol. Chem. 263:12163-12166(1988).
- [9] Valle D., Gaertner J. Nature 361:682-683(1993).
- [10] Aguilar-Bryan L., Nichols C.G., Wechsler S.W., Clement J.P. IV, Boyd A.E. III, Gonzalez G., Herrera-Sosa H., Nguy K., Bryan J., Nelson D.A. Science 268:423-426(1995).

4. (ACBP)

Acyl-CoA-binding protein signature

- 25 Acyl-CoA-binding protein (ACBP) is a small (10 Kd) protein that binds medium- and long-chain acyl-CoA esters with very high affinity and may function as an intracellular carrier of acyl-CoA esters [1]. ACBP is also known as diazepam binding inhibitor (DBI) or endozepine (EP) because of its ability to displace diazepam from the benzodiazepine (BZD) recognition site located on the GABA type A receptor. It is therefore possible that this protein also acts as
- 30 a neuropeptide to modulate the action of the GABA receptor [2].ACBP is a highly conserved protein of about 90 residues that has been so far found in vertebrates, insects and yeast. ACBP is also related to the N-terminal section of a probable transmembrane protein of

unknown function which has been found in mammals. As a signature pattern, the region that corresponds to residues 19 to 37 in mammalian ACBP was selected.

Consensus pattern: P-[STA]-x-[DEN]-x-[LIVMF]-x(2)-[LIVMFY]-Y-[GSTA]-x-[FY]-K- Q-
[STA](2)-x-G-

[1] Rose T.M., Schultz E.R., Todaro G.J. Proc. Natl. Acad. Sci. U.S.A. 89:11287-11291(1992).

[2] Costa E., Guidotti A. Life Sci. 49:325-344(1991).

5. (AIRS)

AIR synthase related proteins

This family includes Hydrogen expression/formation protein HypE, AIR synthases, FGAM synthase and selenide, water dikinase.

6. (AMP-binding)

Putative AMP-binding domain signature

It has been shown [1 to 5] that a number of prokaryotic and eukaryotic enzymes which all probably act via an ATP-dependent covalent binding of AMP to their substrate, share a region of sequence similarity. These enzymes are: - Insects luciferase (luciferin 4-monooxygenase). Luciferase produces light by catalyzing the oxidation of luciferin in presence of ATP and molecular oxygen. - Alpha-aminoadipate reductase from yeast (gene LYS2). This enzyme catalyzes the activation of alpha-aminoadipate by ATP-dependent adenylation and the reduction of activated alpha-aminoadipate by NADPH. - Acetate--CoA ligase (acetyl-CoA synthetase), an enzyme that catalyzes the formation of acetyl-CoA from acetate and CoA. - Long-chain-fatty-acid--CoA ligase, an enzyme that activates long-chain fatty acids for both the synthesis of cellular lipids and their degradation via beta-oxidation. - 4-coumarate--CoA ligase (4CL), a plant enzyme that catalyzes the formation of 4-coumarate-CoA from 4-coumarate and coenzyme A; the branchpoint reactions between

general phenylpropanoid metabolism and pathways leading to various specific end products. - O-succinylbenzoic acid--CoA ligase (OSB-CoA synthetase) (gene *menE*) [6], a bacterial enzyme involved in the biosynthesis of menaquinone (vitamin K2). - 4-Chlorobenzoate--CoA ligase (EC 6.2.1.-) (4-CBA--CoA ligase) [7], a *Pseudomonas* enzyme involved in the

5 degradation of 4-CBA. - Indoleacetate--lysine ligase (IAA-lysine synthetase) [8], an enzyme from *Pseudomonas syringae* that converts indoleacetate to IAA-lysine. - Bile acid-CoA ligase (gene *baiB*) from *Eubacterium* strain VPI 12708 [4]. This enzyme catalyzes the ATP-dependent formation of a variety of C-24 bile acid-CoA. - Crotonobetaine/carnitine-CoA

10 ligase (EC 6.3.2.-) from *Escherichia coli* (gene *caiC*). - L-(alpha-amino adipyl)-L-cysteinyl-D-valine synthetase (ACV synthetase) from various fungi (gene *acvA* or *pcbAB*). This enzyme catalyzes the first step in the biosynthesis of penicillin and cephalosporin, the formation of ACV from the constituent amino acids. The amino acids seem to be activated by adenylation. It is a protein of around 3700 amino acids that contains three related domains of about 1000

15 amino acids. - Gramicidin S synthetase I (gene *grsA*) from *Bacillus brevis*. This enzyme catalyzes the first step in the biosynthesis of the cyclic antibiotic gramicidin S, the ATP-dependent racemization of phenylalanine - Tyrocidine synthetase I (gene *tycA*) from *Bacillus brevis*. The reaction carried out by *tycA* is identical to that catalyzed by *grsA* - Gramicidin S synthetase II (gene *grsB*) from *Bacillus brevis*. This enzyme is a multifunctional protein that activates and polymerizes proline, valine, ornithine and leucine. *GrsB* consists of four related domains. - Enterobactin synthetase components E (gene *entE*)

20 and F (gene *entF*) from *Escherichia coli*. These two enzymes are involved in the ATP-dependent activation of respectively 2,3-dihydroxybenzoate and serine during enterobactin (enterochelin) biosynthesis. - Cyclic peptide antibiotic surfactin synthase subunits 1, 2 and 3 from *Bacillus subtilis*. Subunits 1 and 2 contains three related domains while subunit 3 only

25 contains a single domain. - HC-toxin synthetase (gene *HTS1*) from *Cochliobolus carbonum*. This enzyme activates the four amino acids (Pro, L-Ala, D-Ala and 2-amino-9,10-epoxy-8-oxodecanoic acid) that make up HC-toxin, a cyclic tetrapeptide. *HTS1* consists of four related domains. There are also some proteins, whose exact function is not yet known, but which are, very probably, also AMP-binding enzymes. These proteins are: - ORA (octapeptide-repeat

30 antigen), a *Plasmodium falciparum* protein whose function is not known but which shows a high degree of similarity with the above proteins. - AngR, a *Vibrio anguillarum* protein. AngR is thought to be a transcriptional activator which modulates the anguibactin (an iron-binding siderophore) biosynthesis gene cluster operon. But it is believed [9], that *angR* is not

a DNA-binding protein, but rather an enzyme involved in the biosynthesis of anguibactin. This conclusion is based on three facts: the presence of the AMP-binding domain; the size of angR (1048 residues), which is far bigger than any bacterial transcriptional protein; and the presence of a probable S-acyl thioesterase immediately downstream of angR. - A

5 hypothetical protein in mmsB 3'region in *Pseudomonas aeruginosa*. - *Escherichia coli* hypothetical protein ydiD. - Yeast hypothetical protein YBR041w. - Yeast hypothetical protein YBR222c. - Yeast hypothetical protein YER147c. All these proteins contain a highly conserved region very rich in glycine, serine, and threonine which is followed by a conserved lysine. A parallel can be drawn between this type of domain and the G-x(4)-G-K-[ST] ATP-
10 /GTP-binding 'P-loop' domain or the protein kinases G-x-G-x(2)-[SG]-x(10,20)-KATP-binding domains.

Consensus pattern: [LIVMFY]-x(2)-[STG]-[STAG]-G-[ST]-[STEI]-[SG]-x-[PASLIVM]-
[KR] In a majority of cases the residue that follows the Lys at the end of the pattern is a Gly.

[1] Toh H. Protein Seq. Data Anal. 4:111-117(1991).

[2] Smith D.J., Earl A.J., Turner G. EMBO J. 9:2743-2750(1990).

[3] Schroeder J. Nucleic Acids Res. 17:460-460(1989).

[4] Mallonee D.H., Adams J.L., Hylemon P.B. J. Bacteriol. 174:2065-2071(1992).

[5] Turgay K., Krause M., Marahiel M.A. Mol. Microbiol. 6:529-546(1992).

[6] Driscoll J.R., Taber H.W. J. Bacteriol. 174:5063-5071(1992).

[7] Babbitt P.C., Kenyon G.L., Matin B.M., Charest H., Sylvestre M., Scholten J.D., Chang K.-H., Liang P.-H., Dunaway-Mariano D. Biochemistry 31:5594-5604(1992).

[8] Farrell D.H., Mikesell P., Actis L.A., Crosa J.H. Gene 86:45-51(1990).

7. AP2 domain

This 60 amino acid residue domain can bind to DNA [1]. This domain is plant specific.

30 Members of this family are suggested to be related to pyridoxal phosphate-binding domains such as found in aminotran_2 [3]. AP2 domains are also described in Jofuku et al., co-pending U.S. Patent applications 08/700,152, 08/879,827, 08/912,272, 09/026,039.

- [1] Ohme-takagi M, Shinshi H; Plant Cell 1995;7:173-182.
- [2] Weigel D; Plant Cell 1995;7:388-389.
- [3] Mushegian AR, Koonin EV; Genetics 1996;144:817-828.

5

8. ARID

The ARID domain is an AT-Rich Interaction domain sharing structural homology to DNA replication and repair nucleases and polymerases.

- 10 [1] Herrscher RF, Kaplan MH, Lelsz DL, Das C, Scheuermann R, Tucker PW; Genes Dev 1995;9:3067-3082.
- [2] Yuan YC, Whitson RH, Liu Q, Itakura K, Chen Y; Nat Struct Biol 1998;5:959-964.

15

9. (ATP synt)

ATP synthase gamma subunit signature

20

ATP synthase (proton-translocating ATPase) (EC 3.6.1.34) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria, and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), and a catalytic core, called coupling factor CF(1). The former acts as a proton channel; the latter is composed of five subunits, alpha, beta, gamma, delta and epsilon. Subunit gamma is believed to be important in regulating ATPase activity and the flow of protons through the CF(0) complex. The best conserved region of the gamma subunit [3] is its C-terminus which seems to be essential for assembly and catalysis. As a signature pattern to detect ATPase gamma subunits, a 14 residue conserved segment where the last amino acid is found one to three residues from the C-terminal extremity was used.

25

Consensus pattern: [IV]-T-x-E-x(2)-[DE]-x(3)-G-A-x-[SAKR]- Note: Pea chloroplast gamma and two Bacillus species gamma subunits are not detected by this motif.

30

- [1] Futai M., Noumi T., Maeda M. Annu. Rev. Biochem. 58:111-136(1989).
- [2] Senior A.E. Physiol. Rev. 68:177-231(1988).

[3] Miki J., Maeda M., Mukohata Y., Futai M. FEBS Lett. 232:221-226(1988).

10. (ATP Synt A)

5 Synthase a subunit signature

ATP synthase (proton-translocating ATPase) (EC 3.6.1.34) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria, and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric

10 transmembrane sector, called CF(0), which acts as a proton channel, and a catalytic core, termed coupling factor CF(1). The CF(0) a subunit, also called protein 6, is a key component of the proton channel; it may play a direct role in translocating protons across the membrane. It is a highly hydrophobic protein that has been predicted to contain 8 transmembrane regions [3]. Sequence comparison of a subunits from all available sources reveals very few conserved

15 regions. The best conserved region is located in what is predicted to be the fifth transmembrane domain. This region contains three perfectly conserved residues: an arginine, a leucine and an asparagine. Mutagenesis experiments of ATPase activity. This region was selected as a signature pattern.

20 Consensus pattern: [STAGN]-x-[STAG]-[LIVMF]-R-L-x-[SAGV]-N-[LIVMT] [R is important for proton translocation]

[1] Futai M., Noumi T., Maeda M. Annu. Rev. Biochem. 58:111-136(1989).

[2] Senior A.E. Physiol. Rev. 68:177-231(1988).

25 [3] Lewis M.L., Chang J.A., Simoni R.D. J. Biol. Chem. 265:10541-10550(1990).

[4] Cain B.D., Simoni R.D. J. Biol. Chem. 264:3292-3300(1989).

11. ATP synthase B

30 Part of the CF(0) (base unit) of the ATP synthase. The base unit is thought to translocate protons through membrane (inner membrane in mitochondria, thylakoid membrane in plants, cytoplasmic membrane in bacteria). The B subunits are thought to interact with the stalk of the CF(1) subunits.

12. (ATP synt C)

ATP synthase c subunit signature

5

10

15

20

ATP synthase (proton-translocating ATPase) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria, and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), which acts as a proton channel, and a catalytic core, termed coupling factor CF(1). The CF(0) c subunit (also called protein 9, proteolipid, or subunit III) [3,4] is a highly hydrophobic protein of about 8 Kd which has been implicated in the proton-conducting activity of ATPase. Structurally subunit c consists of two long terminal hydrophobic regions, which probably span the membrane, and a central hydrophilic region. N,N'-dicyclohexylcarbodiimide (DCCD) can bind covalently to subunit c and thereby abolish the ATPase activity. DCCD binds to a specific glutamate or aspartate residue which is located in the middle of the second hydrophobic region near the C-terminus of the protein. A signature pattern which includes the DCCD-binding residue was derived.

Consensus pattern: [GSTA]-R-[NQ]-P-x(10)-[LIVMFYW](2)-x(3)-[LIVMFYW]-x-[DE] [D or E binds DCCD]

25

[1] Futai M., Noumi T., Maeda M. Annu. Rev. Biochem. 58:111-136(1989).

[2] Senior A.E. Physiol. Rev. 68:177-231(1988).

[3] Ivaschenko A.T., Karpenyuk T.A., Ponomarenko S.V. Biokhimiia 56:406-419(1991).

[4] Recipon H., Perasso R., Adoutte A., Quetier F. J. Mol. Evol. 34:292-303(1992).

13. (ATP synt DE)

ATP synthase, Delta/Epsilon chain

30

Part of the ATP synthase CF(1). These subunits are part of the head unit of the ATP synthase. The subunits are called delta and epsilon in human and metazoan species but in bacterial

species the delta (D) subunit is the equivalent to the Oligomycin sensitive subunit (OSCP) in metazoans.

14. (ATP synt ab)

ATP synthase alpha and beta subunits signature

ATP synthase (proton-translocating ATPase) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria, and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), and a catalytic core, called coupling factor CF(1). The former acts as a proton channel; the latter is composed of five subunits, alpha, beta, gamma, delta and epsilon. The sequences of subunits alpha and beta are related and both contain a nucleotide-binding site for ATP and ADP. The beta chain has catalytic activity, while the alpha chain is a regulatory subunit. Vacuolar ATPases [3] (V-ATPases) are responsible for acidifying a variety of intracellular compartments in eukaryotic cells. Like F-ATPases, they are oligomeric complexes of a transmembrane and a catalytic sector. The sequence of the largest subunit of the catalytic sector (70 Kd) is related to that of F-ATPase beta subunit, while a 60 Kd subunit, from the same sector, is related to the F-ATPases alpha subunit [4]. Archaeobacterial membrane-associated ATPases are composed of three subunits. The alpha chain is related to F-ATPases beta chain and the beta chain is related to F-ATPases alpha chain [4]. A protein highly similar to F-ATPase beta subunits is found [5] in some bacterial apparatus involved in a specialized protein export pathway that proceeds without signal peptide cleavage. This protein is known as flhI in *Bacillus* and *Salmonella*, Spa47 (mxlB) in *Shigella flexneri*, HrpB6 in *Xanthomonas campestris* and yscN in *Yersinia* virulence plasmids. To detect these ATPase subunits, a segment of ten amino-acid residues, containing two conserved serines, as a signature pattern was selected. The first serine seems to be important for catalysis - in the ATPase alpha chain at least - as its mutagenesis causes catalytic impairment.

Consensus pattern: P-[SAP]-[LIV]-[DNH]-x(3)-S-x-S [The first S is a putative active site residue]

[1] Futai M., Noumi T., Maeda M. *Annu. Rev. Biochem.* 58:111-136(1989).

[2] Senior A.E. *Physiol. Rev.* 68:177-231(1988).

[3] Nelson N. J. *Bioenerg. Biomembr.* 21:553-571(1989).

[4] Gogarten J.P., Kibak H., Dittrich P., Taiz L., Bowman E.J., Bowman B.J., Manolson M.F., Poole R.J., Date T., Oshima T., Konishi J., Denda K., Yoshida M. *Proc. Natl. Acad. Sci. U.S.A.* 86:6661-6665(1989).

[5] Dreyfus G., Williams A.W., Kawagishi I., MacNab R.M. *J. Bacteriol.* 175:3131-3138(1993).

15. (ATP synt ab C)

ATP synthase ab C terminal.

Number of members: 190

[1] Abrahams JP, Leslie AG, Lutter R, Walker JE; "Structure at 2.8 Å resolution of F1-ATPase from bovine heart mitochondria." *Nature* 1994;370:621-628.

16. (A deaminase)

Adenosine and AMP deaminase signature

Adenosine deaminase catalyzes the hydrolytic deamination of adenosine into inosine. AMP deaminase catalyzes the hydrolytic deamination of AMP into IMP. It has been shown [1] that these two types of enzymes share three regions of sequence similarities; these regions are centered on residues which are proposed to play an important role in the catalytic mechanism of these two enzymes. One of these regions, containing two conserved aspartic acid residues that are potential active site residues was selected.

Consensus pattern: [SA]-[LIVM]-[NGS]-[STA]-D-D-P [The two D's are putative active site residues]

[1] Chang Z., Nygaard P., Chinault A.C., Kellems R.E. *Biochemistry* 30:2273-2280(1991).

17. (Acetyltransf)

Acetyltransferase (GNAT) family.

- 5 This family contains proteins with N-acetyltransferase functions.

[1] Neuwald AF, Landsman D; Trends Biochem Sci 1997;22:154-155.

10 18. (Aconitase C)

Aconitase family signature

Aconitase (aconitate hydratase) (EC 4.2.1.3) [1] is the enzyme from the tricarboxylic acid cycle that catalyzes the reversible isomerization of citrate and isocitrate. Cis-aconitate is formed as an intermediary product during the course of the reaction. In eukaryotes two isozymes of aconitase are known to exist: one found in the mitochondrial matrix and the other found in the cytoplasm. Aconitase, in its active form, contains a 4Fe-4S iron-sulfur cluster; three cysteine residues have been shown to be ligands of the 4Fe-4S cluster. It has been shown that the aconitase family also contains the following proteins: - Iron-responsive element binding protein (IRE-BP). IRE-BP is a cytosolic protein that binds to iron-responsive elements (IREs). IREs are stem-loop structures found in the 5'UTR of ferritin, and delta aminolevulinic acid synthase mRNAs, and in the 3'UTR of transferrin receptor mRNA. IRE-BP also express aconitase activity. - 3-isopropylmalate dehydratase (EC 4.2.1.33) (isopropylmalate isomerase), the enzyme that catalyzes the second step in the biosynthesis of leucine. - Homoaconitase (EC 4.2.1.36) (homoaconitate hydratase), an enzyme that participates in the alpha-aminoadipate pathway of lysine biosynthesis and that converts cis-homoaconitate into homoisocitric acid. - Escherichia coli protein ybhJ. As a signature for proteins from the aconitase family, two conserved regions that contain the three cysteine ligands of the 4Fe-4S cluster were selected.

30 Consensus pattern: [LIVM]-x(2)-[GSACIVM]-x-[LIV]-[GTIV]-[STP]-C-x(0,1)-T-N-[GSTANI]-x(4)-[LIVMA] [C binds the iron-sulfur center]

Consensus pattern: G-x(2)-[LIVWPQ]-x(3)-[GAC]-C-[GSTAM]-[LIMPTA]-C-[LIMV]-
[GA] [The two C's bind the iron-sulfur center]

[1] Gruer M.J., Artymiuk P.J., Guest J.R. Trends Biochem. Sci. 22:3-6(1997).

5

19. (Acyl-CoA dh)

Acyl-CoA dehydrogenases signatures

10 Acyl-CoA dehydrogenases [1,2,3] are enzymes that catalyze the alpha, beta-dehydrogenation
of acyl-CoA esters and transfer electrons to ETF, the electron transfer protein. Acyl-CoA
dehydrogenases are FAD flavoproteins. This family currently includes: - Five eukaryotic
isozymes that catalyze the first step of the beta-oxidation cycles for fatty acids with various
chain lengths. These are short (SCAD) (EC 1.3.99.2), medium (MCAD) (EC 1.3.99.3), long
15 (LCAD) (EC 1.3.99.13), very-long (VLCAD) and short/branched (SBCAD) chain acyl-CoA
dehydrogenases. These enzymes are located in the mitochondrion. They are all
homotetrameric proteins of about 400 amino acid residues except VLCAD which is a dimer
and which contains, in its mature form, about 600 residues. - Glutaryl-CoA dehydrogenase
(EC 1.3.99.7) (GCDH), which is involved in the catabolism of lysine, hydroxylysine and
tryptophan. - Isovaleryl-CoA dehydrogenase (EC 1.3.99.10) (IVD), involved in the
20 catabolism of leucine. - Acyl-coA dehydrogenases acsA and mmgC from *Bacillus subtilis*. -
Butyryl-CoA dehydrogenase (EC 1.3.99.2) from *Clostridium acetobutylicum*. - *Escherichia*
coli protein caiA [4]. - *Escherichia coli* protein aidB. Two conserved regions were selected as
signature patterns. The first is located in the center of these enzymes, the second in the C-
25 terminal section.

Consensus pattern: [GAC]-[LIVM]-[ST]-E-x(2)-[GSAN]-G-[ST]-D-x(2)-[GSA]

Consensus pattern: [QDE]-x(2)-G-[GS]-x-G-[LIVMFY]-x(2)-[DEN]-x(4)-[KR]-x(3)- [DEN]

30

[1] Tanaka K., Ikeda, Matsubara Y., Hyman D.B. Enzyme 38:91-107(1987).

[2] Matsubara Y., Indo Y., Naito E., Ozasa H., Glassberg R., Vockley J., Ikeda Y., Kraus J.,
Tanaka K. J. Biol. Chem. 264:16321-16331(1989).

[3] Aoyama T., Ueno I., Kamijo T., Hashimoto T. J. Biol. Chem. 269:19088-19094(1994).

[4] Eichler K., Bourgis F., Buchet A., Kleber H.-P., Mandrand-Berthelot M.-A. Mol. Microbiol. 13:775-786(1994).

5

20. (Acyl transf)

Acyl transferase domain

Number of members: 161

10

[1] Serre L, Verbree EC, Dauter Z, Stuitje AR, Derewenda ZS; Medline: [95286570](#) "The Escherichia coli malonyl-CoA:acyl carrier protein transacylase at 1.5-A resolution. Crystal structure of a fatty acid synthase component." J Biol Chem 1995;270:12961-12964.

15

21. Acylphosphatase signatures

Acylphosphatase (EC [3.6.1.7](#)) [1,2] catalyzes the hydrolysis of various acylphosphate carboxyl-phosphate bonds such as carbamyl phosphate, succinylphosphate, 1,3-diphosphoglycerate, etc. The physiological role of this enzyme is not yet clear.

20

Acylphosphatase is a small protein of around 100 amino-acid residues. There are two known isozymes. One seems to be specific to muscular tissues, the other, called 'organ-common type', is found in many different tissues. While acylphosphatase have been so far only characterized in vertebrates, there are a number of bacterial and archebacterial hypothetical proteins that are highly similar to that enzyme and that probably possess the same activity. These proteins are: - Escherichia coli hypothetical protein yccX. - Bacillus subtilis hypothetical protein yflL. - Archaeoglobus fulgidus hypothetical protein AF0818. Two conserved regions were selected as signature patterns. The first is located in the N-terminal section, while the second is found in the central part of the protein sequence.

25

Consensus pattern: [LIV]-x-G-x-V-Q-G-V-x-[FM]-R

30

Consensus pattern: G-[FYW]-[AVC]-[KRQAM]-N-x(3)-G-x-V-x(5)-G

[1] Stefani M., Ramponi G. Life Chem. Rep. 12:271-301(1995).

[2] Stefani M., Taddei N., Ramponi G. Cell. Mol. Life Sci. 53:141-151(1997).

5

22. (Adap comp sub)

Clathrin adaptor complexes medium chain signatures.

10

Clathrin coated vesicles (CCV) mediate intracellular membrane traffic such as receptor mediated endocytosis. In addition to clathrin, the CCV are composed of a number of other components including oligomeric complexes which are known as adaptor or clathrin assembly proteins (AP) complexes [1]. The adaptor complexes are believed to interact with the cytoplasmic tails of membrane proteins, leading to their selection and concentration. In mammals two types of adaptor complexes are known: AP-1 which is associated with the Golgi complex and AP-2 which is associated with the plasma membrane. Both AP-1 and AP-2 are heterotetramers that consist of two large chains - the adaptins - (gamma and beta' in AP-1; alpha and beta in AP-2); a medium chain (AP47 in AP-1; AP50 in AP-2) and a small chain (AP19 in AP-1; AP17 in AP-2). The medium chains of AP-1 and AP-2 are evolutionary related proteins of about 50 Kd. Homologs of AP47 and AP50 have also been found in *Caenorhabditis elegans* (genes unc-101 and ap50) [2] and yeast (gene APM1 or YAP54) [3]. Some more divergent, but clearly evolutionary related proteins have also been found in yeast: APM2 and YBR288c. Two conserved regions were selected as signature patterns, one located in the N-terminal region, the other from the central section of these proteins.

25

Consensus pattern: [IVT]-[GSP]-W-R-x(2,3)-[GAD]-x(2)-[HY]-x(2)-N-x-[LIVMAFY](3)-D-[LIVM]-[LIVMT]-E

Consensus pattern: [LIV]-x-F-I-P-P-x-G-x-[LIVMFY]-x-L-x(2)-Y

30

[1] Pearse B.M., Robinson M.S. Annu. Rev. Cell Biol. 6:151-171(1990).

[2] Lee J., Jongeward G.D., Sternberg P.W. Genes Dev. 8:60-73(1994).

[3] Nakayama Y., Goebel M., O'Brine G.B., Lemmon S., Pingchang C.E., Kirchhausen T. Eur. J. Biochem. 202:569-574(1991).

23. (Adenylsucc synt)

Adenylosuccinate synthetase signatures

5

Adenylosuccinate synthetase (EC 6.3.4.4) [1] plays an important role in purine biosynthesis, by catalyzing the GTP-dependent conversion of IMP and aspartic acid to AMP.

Adenylosuccinate synthetase has been characterized from various sources ranging from *Escherichia coli* (gene *purA*) to vertebrate tissues. Invertebrates, two isozymes are present - one involved in purine biosynthesis and the other in the purine nucleotide cycle. Two conserved regions were selected as signature patterns. The first one is a perfectly conserved octapeptide located in the N-terminal section and which is involved in GTP-binding [2]. The second one includes a lysine residue known [2] to be essential for the enzyme's activity.

10

Consensus pattern: Q-W-G-D-E-G-K-G

Consensus pattern: G-I-[GR]-P-x-Y-x(2)-K-x(2)-R [K is the active site residue]

[1] Wiesmueller L., Wittbrodt J., Noegel A.A., Schleicher M. J. Biol. Chem. 266:2480-2485(1991).

[2] Silva M.M., Poland B.W., Hoffman C.R., Fromm H.J., Honzatko R.B. J. Mol. Biol. 254:431-446(1995).

[3] Bouyoub A., Barbier G., Forterre P., Labedan B. 2.3.CO;2-J. Mol. Biol. 261:144-154(1996).

25

24. (AdoHcyase)

S-adenosyl-L-homocysteine hydrolase signatures

30

S-adenosyl-L-homocysteine hydrolase (EC 3.3.1.1) (AdoHcyase) is an enzyme of the activated methyl cycle, responsible for the reversible hydration of S-adenosyl-L-homocysteine into adenosine and homocysteine. AdoHcyase is a ubiquitous enzyme which binds and requires NAD⁺ as a cofactor. AdoHcyase is a highly conserved protein [1] of about

430 to 470 amino acids. Two highly conserved regions were selected as signature patterns. The first pattern is located in the N-terminal section; the second is derived from aglycine-rich region in the central part of AdoHcyase; a region thought to be involved in NAD-binding.

5 Consensus pattern: [GSA]-[CS]-N-x-[FYLM]-S-[ST]-[QA]-[DEN]-x-[AV]-[AT]-[AD]-[AC]-[LIVMCG]

Consensus pattern: [GA]-[KS]-x(3)-[LIV]-x-G-[FY]-G-x-[VC]-G-[KRL]-G-x-[ASC]

10 [1] Sganga M.W., Aksamit R.R., Cantoni G.L., Bauer C.E. Proc. Natl. Acad. Sci. U.S.A. 89:6328-6332(1992).

25. AhpC/TSA family

15 This family contains proteins related to alkyl hydroperoxide reductaseComment: (AhpC) and thiol specific antioxidant (TSA).

[1] Chae HZ, Robison K, Poole LB, Church G, Storz G, Rhee SG, Proc Natl Acad Sci U S A 1994;91:7017-7021

26. (Aldose epim)

25 Aldose 1-epimerase putative active site Aldose 1-epimerase (EC 5.1.3.3) (mutarotase) is the enzyme responsible for the anomeric interconversion of D-glucose and other aldoses between their alpha- and beta-forms. The sequence of mutarotase from two bacteria, *Acinetobacter calcoaceticus* and *Streptococcus thermophilus* is available [1]. It has also been shown that, on the basis of extensive sequence similarities, a mutarotase domain seem to be present in the C-terminal half of the fungal GAL10 protein which encodes, in the N-terminal

30 part, for UDP-glucose 4-epimerase. The best conserved region in the sequence of mutarotase is centered around a conserved histidine residue which may be involved in the catalytic mechanism.

Consensus pattern: [NS]-x-T-N-H-x-Y-[FW]-N-[LI]

[1] Poolman B., Royer T.J., Mainzer S.E., Schmidt B.F. J. Bacteriol. 172:4037-4047(1990).

5

27. (AlkA DNA repair)

Alkylbase DNA glycosidases alkA family signature

10

Alkylbase DNA glycosidases [1] are DNA repair enzymes that hydrolyzes the deoxyribose N-glycosidic bond to excise various alkylated bases from a damaged DNA polymer. In Escherichia coli there are two alkylbase DNA glycosidases: one (gene tag) which is constitutively expressed and which is specific for the removal of 3-methyladenine (EC 3.2.2.20), and one (gene alkA) which is induced during adaptation to alkylation and which can remove a variety of alkylation products (EC 3.2.2.21). Tag and alkA do not share any region of sequence similarity. In yeast there is an alkylbase DNA glycosidase (gene MAG1) [2,3], which can remove 3-methyladenine or 7-methyladenine and which is structurally related to alkA. MAG and alkA are both proteins of about 300 amino acid residues. While the C- and N-terminal ends appear to be unrelated, there is a central region of about 130 residues which is well conserved. A portion of this region has been selected as a signature pattern .

15

20

Consensus pattern: G-I-G-x-W-[ST]-[AV]-x-[LIVMFY](2)-x-[LIVM]-x(8)-[MF]-x(2)-[ED]-D

25

[1] Lindahl T., Sedgwick B. Annu. Rev. Biochem. 57:133-157(1988).

[2] Berdal K.G., Bjoras M., Bjelland S., Seeberg E.C. EMBO J. 9:4563-4568(1990).

[3] Chen J., Derfler B., Samson L. EMBO J. 9:4569-4575(1990).

30

28. Ammonium transporters signature

A number of proteins involved in the transport of ammonium ions across a membrane as well as some yet uncharacterized proteins have been shown [1,2] to be evolutionary related. These

proteins are: - Yeast ammonium transporters MEP1, MEP2 and MEP3. - Arabidopsis
thaliana high affinity ammonium transporter (gene AMT1). - *Corynebacterium glutamicum*
 ammonium and methylammonium transport system. - *Escherichia coli* putative ammonium
 transporter amtB. - *Bacillus subtilis* nrgA. - *Mycobacterium tuberculosis* hypothetical
 5 protein MtCY338.09c. - *Synechocystis* strain PCC 6803 hypothetical proteins slt0108,
 slt0537 and slt1017. - *Methanococcus jannaschii* hypothetical proteins MJ0058 and MJ1343.
 - *Caenorhabditis elegans* hypothetical proteins C05E11.4, F49E11.3 and M195.3. As
 expected by their transport function, these proteins are highly hydrophobic and seem to
 contain from 10 to 12 transmembrane domains. The best conserved region seems to be
 10 located in the fifth (or sixth) transmembrane region and is used as a signature pattern.

Consensus pattern: D-[FYWS]-A-G-[GSC]-x(2)-[IV]-x(3)-[SAG](2)-x(2)-[SAG]- [LIVMF]-
 x(3)-[LIVMFYWA](2)-x-[GK]-x-R

- 15 [1] Ninnemann O., Janniaux J.-C., Frommer W.B. EMBO J. 13:3464-3471(1994).
 [2] Siewe R.M., Weil B., Burkovski A., Eikmanns B.J., Eikmanns M., Kraemer R. J. Biol.
 Chem. 271:5398-5403(1996).
 [3] Saier M.H. Jr. Adv. Microbiol. Physiol. 40:81-136(1998).

20 29. (Arch_histone)

CBF/NF-Y subunits signatures

Diverse DNA binding proteins are known to bind the CCAAT box, a common cis-acting
 25 element found in the promoter and enhancer regions of a large number of genes in
 eukaryotes. Amongst these proteins is one known as the CCAAT-binding factor (CBF) or
 NF-Y [1]. CBF is a heteromeric transcription factor that consists of two different components
 both needed for DNA-binding. The HAP protein complex of yeast binds to the upstream
 activation site of cytochrome C iso-1 gene (CYC1) as well as other genes involved in
 30 mitochondrial electron transport and activates their expression. It also recognizes the
 sequence CCAAT and is structurally and evolutionary related to CBF. The first subunit of
 CBF, known as CBF-A or NF-YB in vertebrates, HAP3 in budding yeast and as php3 in
 fission yeast, is a protein of 116 to 210 amino-acid residues which contains a highly

conserved central domain of about 90 residues. This domain seems to be involved in DNA-binding; a signature pattern had been developed from its central part. The second subunit of CBF, known as CBF-B or NF-YA in vertebrates, HAP2 in budding yeast and php2 in fission yeast, is a protein of 265 to 350 amino-acid residues which contains a highly conserved region of about 60 residues. This region, called the 'essential core' [2], seems to consist of two subdomains: an N-terminal subunit-association domain and a C-terminal DNA recognition domain. A signature pattern has been developed from a section of the subunit-association domain.

Consensus pattern: C-V-S-E-x-I-S-F-[LIVM]-T-[SG]-E-A-[SC]-[DE]-[KRQ]-C-

Consensus pattern: Y-V-N-A-K-Q-Y-x-R-I-L-K-R-R-x-A-R-A-K-L-E-

[1] Li X.-Y., Mantovani R., Hooft van Huijsduijnen R., Andre I., Benoist C., Mathis D. Nucleic Acids Res. 20:1087-1091(1992).

[2] Olesen J.T., Fikes J.D., Guarente L. Mol. Cell. Biol. 11:611-619(1991).

30. Argininosuccinate synthase signatures

Argininosuccinate synthase (EC 6.3.4.5) (AS) is a urea cycle enzyme that catalyzes the penultimate step in arginine biosynthesis: the ATP-dependent ligation of citrulline to aspartate to form argininosuccinate, AMP and pyrophosphate [1,2]. In humans, a defect in the AS gene causes citrullinemia, a genetic disease characterized by severe vomiting spells and mental retardation. AS is a homotetrameric enzyme of chains of about 400 amino-acid residues. An arginine seems to be important for the enzyme's catalytic mechanism. The sequences of AS from various prokaryotes, archaeobacteria and eukaryotes show significant similarity. Two signature patterns have been selected for AS. The first is a highly conserved stretch of nine residues located in the N-terminal extremity of these enzymes, the second is derived from a conserved region which contains one of the conserved arginine residues.

Consensus pattern: [AS]-[FY]-S-G-G-[LV]-D-T-[ST]-

Consensus pattern: G-x-T-x-K-G-N-D-x(2)-R-F-

[1] van Vliet F., Crabeel M., Boyen A., Tricot C., Stalon V., Falmagne P., Nakamura Y.,
Baumberg S., Glansdorff N. Gene 95:99-104(1990).

5 [2] Morris C.J., Reeve J.N. J. Bacteriol. 170:3125-3130(1988).

31. Armadillo/beta-catenin-like repeats

Approx. 40 amino acid repeat. Tandem repeats form super-helix of helices that is proposed to
10 mediate interaction of beta-catenin with its ligands. CAUTION: This family does not contain
all known armadillo repeats.

[1] Huber AH, Nelson WJ, Weis WI, Cell 1997;90:871-882.

[2] Gumbiner BM, Curr Opin Cell Biol 1995;7:634-640.

15 [3] Cavallo R, Rubenstein D, Peifer M, Curr Opin Genet Dev 1997;7:459-466.

[4] Su LK, Vogelstein B, Kinzler KW, Science 1993;262:1734-1737.

[5] Masiarz FR, Munemitsu S, Polakis P Science 1993;262:1731-1734

[6] Peifer M, Wieschaus E, Cell 1990;63:1167-1176.

32. (Asn Synthase)

Asparagine synthase

This family is always found associated with GATase_2. Members of this family catalyse the
25 conversion of aspartate to asparagine.

33. Asparaginase_2

Asparaginase 12 members

34. (Aspartyl tRNA N)

Aminoacyl-transfer RNA synthetases class-II signatures

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA

synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is different to the Rossmann fold observed for the class I synthetases [7]. Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns have been derived from two of these regions.

Consensus pattern: [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]

Consensus pattern: [GSTALVF]-{DENQHRKP}-[GSTA]-[LIVMF]-[DE]-R-[LIVMF]-x-[LIVMSTAG]-[LIVMFY]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).

[2] Delarue M., Moras D. BioEssays 15:675-687(1993).

[3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).

[4] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

[5] Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).

[6] Cusack S. Biochimie 75:1077-1081(1993).

[7] Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-255(1990).

[8] Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).

35. (ArfGap) Putative GTP-ase activating protein for Arf. Putative zinc fingers with GTPase activating proteins (GAPs) towards the small GTPase, Arf. The GAP of ARD1 stimulates

GTPase hydrolysis for ARD1 but not ARFs. Number of members: 34

[1]Medline: 96324970. Identification and cloning of centaurin-alpha. A novel phosphatidylinositol 3,4,5-trisphosphate-binding protein from rat brain. Hammonds-Odie LP, Jackson TR, Profit AA, Blader IJ, Turck CW, Prestwich GD, Theibert AB; J Biol Chem 1996;271:18859-18868.

[2]Medline: 97296423. A target of phosphatidylinositol 3,4,5-trisphosphate with a zinc finger motif similar to that of the ADP-ribosylation -factor GTPase-activating protein and two pleckstrin homology domains. Tanaka K, Imajoh-Ohmi S, Sawada T, Shirai R, Hashimoto Y, Iwasaki S, Kaibuchi K, Kanaho Y, Shirai T, Terada Y, Kimura K, Nagata S, Fukui Y; Eur J Biochem 1997;245:512-519.

[3] 98112795. Molecular characterization of the GTPase-activating domain of ADP-ribosylation factor domain protein 1 (ARD1). Vitale N, Moss J, Vaughan M; J Biol Chem 1998;273:2553-2560.

36. Apolipoprotein. Apolipoprotein A1/A4/E family. This family includes: Swiss:P02647 Apolipoprotein A-I. Swiss:P06727 Apolipoprotein A-IV. Swiss:P02649 Apolipoprotein E. These proteins contain several 22 residue repeats which form a pair of alpha helices. Number of members: 42

[1]Medline: 91289138. Three-dimensional structure of the LDL receptor-binding domain of human apolipoprotein E. Wilson C, Wardell MR, Weisgraber KH, Mahley RW, Agard DA; Science 1991;252:1817-1822.

37. Amino acid permeases signature

Amino acid permeases are integral membrane proteins involved in the transport of amino acids into the cell. A number of such proteins have been found to be evolutionary related

[1,2,3]. These proteins are: - Yeast general amino acid permeases (genes GAP1, AGP2 and AGP3). - Yeast basic amino acid permease (gene ALP1). - Yeast Leu/Val/Ile permease (gene BAP2). - Yeast arginine permease (gene CAN1). - Yeast dicarboxylic amino acid permease (gene DIP5). - Yeast asparagine/glutamine permease (gene AGP1). - Yeast glutamine

permease (gene GNP1). - Yeast histidine permease (gene HIP1). - Yeast lysine permease (gene LYP1). - Yeast proline permease (gene PUT4). - Yeast valine and tyrosine permease (gene VAL1/TAT1). - Yeast tryptophan permease (gene TAT2/SCM2). - Yeast choline transport protein (gene HNM1/CTR1). - Yeast GABA permease (gene UGA4). - Yeast hypothetical protein YKL174c. - Fission yeast protein isp5. - Fission yeast hypothetical protein SpAC8A4.11 - Fission yeast hypothetical protein SpAC11D3.08c. - *Emericella nidulans* proline transport protein (gene prnB). - *Trichoderma harzianum* amino acid permease INDA1. - *Salmonella typhimurium* L-asparagine permease (gene ansP). - *Escherichia coli* aromatic amino acid transport protein (gene aroP). - *Escherichia coli* D-serine/D-alanine/glycine transporter (gene cycA). - *Escherichia coli* GABA permease (gene gabP). - *Escherichia coli* lysine-specific permease (gene lysP). - *Escherichia coli* phenylalanine-specific permease (gene pheP). - *Salmonella typhimurium* proline-specific permease (gene proY). - *Escherichia coli* and *Klebsiella pneumoniae* hypothetical protein yeeF. - *Escherichia coli* and *Salmonella typhimurium* hypothetical protein yifK. - *Bacillus subtilis* permeases rocC and rocE which probably transports arginine or ornithine. These proteins seem to contain up to 12 transmembrane segments. As a signature for this family of proteins, the best conserved region which is located in the second transmembrane segment has been selected.

Consensus pattern: [STAGC]-G-[PAG]-x(2,3)-[LIVMFYWA](2)-x-[LIVMFYW]-x-[LIVFWSTAGC](2)-[STAGC]-x(3)-[LIVMFYWT]-x-[LIVMST]-x(3)-[LIVMCTA]-[GA]-E-x(5)-[PSAL]-

[1] Weber E., Chevalier M.R., Jund R. J. Mol. Evol. 27:341-350(1988).

[2] Vandenbol M., Jauniaux J.-C., Grenson M. Gene 83:153-159(1989).

[3] Reizer J., Finley K., Kakuda D., McLeod C.L., Reizer A., Saier M.H. Jr. Protein Sci. 2:20-30(1993).

38. aak kinase (1) Glutamate 5-kinase signature

Glutamate 5-kinase (EC 2.7.2.11) (gamma-glutamyl kinase) (GK) is the enzyme that catalyzes the first step in the biosynthesis of proline from glutamate, the ATP-dependent phosphorylation of L-glutamate into L-glutamate 5-phosphate. In eubacteria (gene proB) and

yeast [1] (gene PRO1), GK is a monofunctional protein, while in plants and mammals, it is a bifunctional enzyme (P5CS) [2] that consists of two domains: a N-terminal GK domain and a C-terminal gamma-glutamyl phosphate reductase domain (EC 1.2.1.41) (see

<PDOC00940>). As a signature pattern, a highly conserved glycine- and alanine-rich region located in the central section of these enzymes has been selected. Yeast hypothetical protein YHR033w is highly similar to GK.

Consensus pattern: [GSTN]-x(2)-G-x-G-[GC]-[IM]-x-[STA]-K-[LIVM]-x-[SA]-[TCA]-x(2)-[GALV]-x(3)-G-

[1] Li W., Brandriss M.C. J. Bacteriol. 174:4148-4156(1992).

[2] Hu C.-A.A., Delauney A.J., Verma D.P.S. Proc. Natl. Acad. Sci. U.S.A. 89:9354-9358(1992).

aakinase (2) Aspartokinase signature

Aspartokinase (EC 2.7.2.4) (AK) [1] catalyzes the phosphorylation of aspartate. The product of this reaction can then be used in the biosynthesis of lysine or in the pathway leading to homoserine, which participates in the biosynthesis of threonine, isoleucine and methionine. In *Escherichia coli*, there are three different isozymes which differ in their sensitivity to repression and inhibition by Lys, Met and Thr. AK1 (gene thrA) and AK2 (gene metL) are bifunctional enzymes which both consist of an N-terminal AK domain and a C-terminal homoserine dehydrogenase domain. AK1 is involved in threonine biosynthesis and AK2, in that of methionine. The third isozyme, AK3 (gene lysC), is monofunctional and involved in lysine synthesis. In yeast, there is a single isozyme of AK (gene HOM3). As a signature pattern for AK, a conserved region located in the N-terminal extremity has been selected.

Consensus pattern: [LIVM]-x-K-[FY]-G-G-[ST]-[SC]-[LIVM]-

[1] Rafalski J.A., Falco S.C. J. Biol. Chem. 263:2146-2151(1988).

aakinase (3) Gamma-glutamyl phosphate reductase signature

Gamma-glutamyl phosphate reductase (EC 1.2.1.41) (GPR) is the enzyme that catalyzes the second step in the biosynthesis of proline from glutamate, the NADP-dependent reduction of

L-glutamate 5-phosphate into L-glutamate 5-semialdehyde and phosphate. In eubacteria (gene proA) and yeast [1] (gene PRO2), GPR is a monofunctional protein, while in plants and mammals, it is a bifunctional enzyme (P5CS) [2] that consists of two domains: a N-terminal glutamate 5-kinase domain (EC 2.7.2.11) (see <PDOC00701>) and a C-terminal GPR domain. As a signature pattern, a conserved region that contains two histidine residues has been selected. This region is located in the last third of GPR.

Consensus pattern: V-x(5)-A-[LIV]-x-H-I-x(2)-[HY]-[GS]-[ST]-x-H-[ST]-[DE]-x-I-

[1] Pearson B.M., Hernando Y., Payne J., Wolf S.S., Kalogeropoulos A., Schweizer M. Yeast 12:1021-1031(1996).

[2] Hu C.-A.A., Delauney A.J., Verma D.P.S. Proc. Natl. Acad. Sci. U.S.A. 89:9354-9358(1992).

39. (abhydrolase) alpha/beta hydrolase fold. This catalytic domain is found in a very wide range of enzymes.

[1] Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, Harel M, Remington SJ, Silman I, Schrag J, Sussman JL, Verschueren KHG, Goldman A, Protein Eng 1992;5:197-211.

40. (Acid phosphat) Histidine acid phosphatases signatures

Acid phosphatases (EC 3.1.3.2) are a heterogeneous group of proteins that hydrolyze phosphate esters, optimally at low pH. It has been shown [1] that a number of acid phosphatases, from both prokaryotes and eukaryotes, share two regions of sequence similarity, each centered around a conserved histidine residue. These two histidines seem to be involved in the enzymes' catalytic mechanism [2,3]. The first histidine is located in the N-terminal section and forms a phosphohistidine intermediate while the second is located in the C-terminal section and possibly acts as proton donor. Enzymes belonging to this family are called 'histidine acid phosphatases' and are listed below:

- Escherichia coli pH 2.5 acid phosphatase (gene appA).

□

- Escherichia coli glucose-1-phosphatase (EC 3.1.3.10) (gene agp).

- Yeast constitutive and repressible acid phosphatases (genes PHO3 and PHO5).

5 - Fission yeast acid phosphatase (gene pho1).

- Aspergillus phytases A and B (EC 3.1.3.8) (gene phyA and phyB).

- Mammalian lysosomal acid phosphatase.

- Mammalian prostatic acid phosphatase.

- Caenorhabditis elegans hypothetical proteins B0361.7, C05C10.1, C05C10.4

10 and F26C11.1.

Consensus pattern[LIVM]-x(2)-[LIVMA]-x(2)-[LIVM]-x-R-H-[GN]-x-R-x-[PAS] [H is the phosphohistidine residue]

□

15 Consensus pattern[LIVMF]-x-[LIVMFAG]-x(2)-[STAGI]-H-D-[STANQ]-x-[LIVM]-x(2)-[LIVMFY]-x(2)-[STA] [H is an active site residue] Sequences known to belong to this class detected by the patternALL, except for rat prostatic acid phosphatase which seems to have Tyr instead of the active site His

20 [1] van Etten R.L., Davidson R., Stevis P.E., MacArthur H., Moore D.L. J. Biol. Chem. 266:2313-2319(1991).

[2] Ostanin K., Harms E.H., Stevis P.E., Kuciel R., Zhou M.-M., van Etten R.L. J. Biol. Chem. 267:22830-22836(1992).

[3] Schneider G., Lindqvist Y., Vihko P. EMBO J. 12:2609-2615(1993).

25

41. Aconitase family signatures

Aconitase (aconitate hydratase) (EC 4.2.1.3) [1] is the enzyme from the tricarboxylic acid cycle that catalyzes the reversible isomerization of citrate and isocitrate. Cis-aconitate is
30 formed as an intermediary product during the course of the reaction. In eukaryotes two isozymes of aconitase are known to exist: one found in the mitochondrial matrix and the other found in the cytoplasm. Aconitase, in its active form, contains a 4Fe-4S iron-sulfur cluster; three cysteine residues have been shown to be ligands of the 4Fe-4S cluster. It has

been shown that the aconitase family also contains the following proteins: - Iron-responsive element binding protein (IRE-BP). IRE-BP is a cytosolic protein that binds to iron-responsive elements (IREs). IREs are stem-loop structures found in the 5'UTR of ferritin, and delta aminolevulinic acid synthase mRNAs, and in the 3'UTR of transferrin receptor mRNA. IRE-BP also express aconitase activity. - 3-isopropylmalate dehydratase (EC 4.2.1.33) (isopropylmalate isomerase), the enzyme that catalyzes the second step in the biosynthesis of leucine. - Homoaconitase (EC 4.2.1.36) (homoaconitate hydratase), an enzyme that participates in the alpha-aminoadipate pathway of lysine biosynthesis and that converts cis-homoaconitate into homoisocitric acid. - Esherichia coli protein ybhJ

Consensus pattern: [LIVM]-x(2)-[GSACIVM]-x-[LIV]-[GTIV]-[STP]-C-x(0,1)-T-N-[GSTANI]-x(4)-[LIVMA] [C binds the iron-sulfur center]

Consensus pattern: G-x(2)-[LIVWPQ]-x(3)-[GAC]-C-[GSTAM]-[LIMPTA]-C-[LIMV]-[GA] [The two C's bind the iron-sulfur center]-

[1] Gruer M.J., Artymiuk P.J., Guest J.R. Trends Biochem. Sci. 22:3-6(1997).

42. Actins signatures

Actins [1 to 4] are highly conserved contractile proteins that are present in all eukaryotic cells. In vertebrates there are three groups of actin isoforms: alpha, beta and gamma. The alpha actins are found in muscle tissues and are a major constituent of the contractile apparatus. The beta and gamma actins co-exists in most cell types as components of the cytoskeleton and as mediators of internal cell motility. In plants [5] there are many isoforms which are probably involved in a variety of functions such as cytoplasmic streaming, cell shape determination, tip growth, graviperception, cell wall deposition, etc. Actin exists either in a monomeric form (G-actin) or in a polymerized form (F-actin). Each actin monomer can bind a molecule of ATP; when polymerization occurs, the ATP is hydrolyzed. Actin is a protein of from 374 to 379 amino acid residues. The structure of actin has been highly conserved in the course of evolution. Recently some divergent actin-like proteins have been identified in several species. These proteins are: - Centractin (actin-RPV) from mammals, fungi (yeast ACT5, Neurospora crassa ro-4) and Pneumocystis carinii (actin-II). Centractin seems to be a component of a multi-subunit centrosomal complex involved in microtubule

based vesicle motility. This subfamily is also known as ARP1. - ARP2 subfamily which includes chicken ACTL, yeast ACT2, Drosophila 14D, C.elegans actC. - ARP3 subfamily which includes actin 2 from mammals, Drosophila 66B, yeast ACT4 and fission yeast act2. - ARP4 subfamily which includes yeast ACT3 and Drosophila 13E. Three signature patterns have been developed. The first two are specific to actins and span positions 54 to 64 and 357 to 365. The last signature picks up both actins and the actin-like proteins and corresponds to positions 106 to 118 in actins.

Consensus pattern: [FY]-[LIV]-G-[DE]-E-A-Q-x-[RKQ](2)-G-

Consensus pattern: W-[IV]-[STA]-[RK]-x-[DE]-Y-[DNE]-[DE]-

Consensus pattern: [LM]-[LIVM]-T-E-[GAPQ]-x-[LIVMFYWHQ]-N-[PSTAQ]-x(2)-N-[KR]-

[1] Sheterline P., Clayton J., Sparrow J.C. (In) Actins, 3rd Edition, Academic Press Ltd, London, (1996).

[2] Pollard T.D., Cooper J.A. Annu. Rev. Biochem. 55:987-1036(1986).

[3] Pollard T.D. Curr. Opin. Cell Biol. 1:33-40(1990).

[4] Rubenstein P.A. BioEssays 12:309-315(1990).

[5] Meagher R.B., McLean B.G. Cell Motil. Cytoskeleton 16:164-166(1990).

43. Adenylate kinase signature

Adenylate kinase (EC 2.7.4.3) (AK) [1] is a small monomeric enzyme that catalyzes the reversible transfer of MgATP to AMP ($\text{MgATP} + \text{AMP} = \text{MgADP} + \text{ADP}$). In mammals there are three different isozymes: - AK1 (or myokinase), which is cytosolic. - AK2, which is located in the outer compartment of mitochondria. - AK3 (or GTP:AMP phosphotransferase), which is located in the mitochondrial matrix and which uses MgGTP instead of MgATP. The sequence of AK has also been obtained from different bacterial species and from plants and fungi. Two other enzymes have been found to be evolutionary related to AK. These are: - Yeast uridylate kinase (EC 2.7.4.-) (UK) (gene URA6) [2] which catalyzes the transfer of a phosphate group from ATP to UMP to form UDP and ADP. - Slime mold UMP-CMP kinase (EC 2.7.4.14) [3] which catalyzes the transfer of a phosphate group from ATP to either CMP or UMP to form CDP or UDP and ADP. Several regions of AK family enzymes are well

conserved, including the ATP-binding domains. The most conserved of all regions have been selected as a signature for this type of enzyme. This region includes an aspartic acid residue that is part of the catalytic cleft of the enzyme and that is involved in a salt bridge. It also includes an arginine residue whose modification leads to inactivation of the enzyme

5

Consensus pattern: [LIVMFYW](3)-D-G-[FYI]-P-R-x(3)-[NQ]-

[1] Schulz G.E. Cold Spring Harbor Symp. Quant. Biol. 52:429-439(1987).

[2] Liljelund P., Sanni A., Friesen J.D., Lacroute F. Biochem. Biophys. Res. Commun. 165:464-473(1989).

10

[3] Wiesmueller L., Noegel A.A., Barzu O., Gerisch G., Schleicher M. J. Biol. Chem. 265:6339-6345(1990).

[4] Kath T.H., Schmid R., Schaefer G. Arch. Biochem. Biophys. 307:405-410(1993).

15

44. (adh_short) Short-chain dehydrogenases/reductases family signature. The short-chain dehydrogenases/reductases family (SDR) [1] is a very large family of enzymes, most of which are known to be NAD- or NADP-dependent oxidoreductases. As the first member of this family to be characterized was *Drosophila* alcohol dehydrogenase, this family used to be called [2,3,4]'insect-type', or 'short-chain' alcohol dehydrogenases. Most member of this family are proteins of about 250 to 300 amino acid residues. The proteins currently known to belong to this family are listed below. - Alcohol dehydrogenase (EC 1.1.1.1) from insects such as *Drosophila*. - Acetoin dehydrogenase (EC 1.1.1.5) from *Klebsiella terrigena* (gene budC). - D-beta-hydroxybutyrate dehydrogenase (BDH) (EC 1.1.1.30) from mammals. - Acetoacetyl-CoA reductase (EC 1.1.1.36) from various bacterial species (gene phbB or phaB). - Glucose 1-dehydrogenase (EC 1.1.1.47) from *Bacillus*. - 3-beta-hydroxysteroid dehydrogenase (EC 1.1.1.51) from *Comomonas testosteroni*. - 20-beta-hydroxysteroid dehydrogenase (EC 1.1.1.53) from *Streptomyces hydrogenans*. - Ribitol dehydrogenase (EC 1.1.1.56) (RDH) from *Klebsiella aerogenes*. - Estradiol 17-beta-dehydrogenase (EC 1.1.1.62) from human. - Gluconate 5-dehydrogenase (EC 1.1.1.69) from *Gluconobacter oxydans* (gene gno). - 3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100) from *Escherichia coli* (gene fabG) and from plants. - Retinol dehydrogenase (EC 1.1.1.105) from mammals. - 2-deoxy-d-gluconate 3-dehydrogenase (EC 1.1.1.125) from *Escherichia coli* and *Erwinia chrysanthemi*

25

20

30

- (gene *kduD*). - Sorbitol-6-phosphate 2-dehydrogenase (EC 1.1.1.140) from *Escherichia coli* (gene *gutD*) and from *Klebsiella pneumoniae* (gene *sorD*). - 15-hydroxyprostaglandin dehydrogenase (NAD⁺) (EC 1.1.1.141) from human. - Corticosteroid 11-beta-dehydrogenase (EC 1.1.1.146) (11-DH) from mammals. - 7-alpha-hydroxysteroid dehydrogenase (EC 1.1.1.159) from *Escherichia coli* (gene *hdhA*), *Eubacterium* strain VPI 12708 (gene *baiA*) and from *Clostridium sordellii*. - NADPH-dependent carbonyl reductase (EC 1.1.1.184) from mammals. - Tropinone reductase-I (EC 1.1.1.206) and -II (EC 1.1.1.236) from plants. - N-acylmannosamine 1-dehydrogenase (EC 1.1.1.233) from *Flavobacterium* strain 141-8. - D-arabinitol 2-dehydrogenase (ribulose forming) (EC 1.1.1.250) from fungi. -
- 10 Tetrahydroxynaphthalene reductase (EC 1.1.1.252) from *Magnaporthe grisea*. - Pteridine reductase 1 (EC 1.1.1.253) (gene *PTR1*) from *Leishmania*. - 2,5-dichloro-2,5-cyclohexadiene-1,4-diol dehydrogenase (EC 1.1.-.-) from *Pseudomonas paucimobilis*. - Cis-1,2-dihydroxy-3,4-cyclohexadiene-1-carboxylate dehydrogenase (EC 1.3.1.-) from *Acinetobacter calcoaceticus* (gene *benD*) and *Pseudomonas putida* (gene *xylL*). - Biphenyl-2,3-dihydro-2,3-diol dehydrogenase (EC 1.3.1.-) (gene *bphB*) from various *Pseudomonaceae*. - Cis-toluene dihydrodiol dehydrogenase (EC 1.3.1.-) from *Pseudomonas putida* (gene *todD*). - Cis-benzene glycol dehydrogenase (EC 1.3.1.19) from *Pseudomonas putida* (gene *bnzE*). - 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase (EC 1.3.1.28) from *Escherichia coli* (gene *entA*) and *Bacillus subtilis* (gene *dhbA*). - Dihydropteridine reductase (EC 1.6.99.7) (HDHPR) from mammals. - Lignin degradation enzyme *ligD* from *Pseudomonas paucimobilis*. - Agropine synthesis reductase from *Agrobacterium* plasmids (gene *mas1*). - Versicolorin reductase from *Aspergillus parasiticus* (gene *VER1*). - Putative keto-acyl reductases from *Streptomyces* polyketide biosynthesis operons. - A trifunctional hydratase-dehydrogenase-epimerase from the peroxisomal beta-oxidation system of *Candida tropicalis*.
- 25 This protein contains two tandemly repeated 'short-chain dehydrogenase-type' domain in its N-terminal extremity. - Nodulation protein *nodG* from species of *Azospirillum* and *Rhizobium* which is probably involved in the modification of the nodulation Nod factor fatty acyl chain. - Nitrogen fixation protein *fixR* from *Bradyrhizobium japonicum*. - *Bacillus subtilis* protein *dltE* which is involved in the biosynthesis of D-alanyl-lipoteichoic acid. -
- 30 Human follicular variant translocation protein 1 (FVT1). - Mouse adipocyte protein p27. - Mouse protein Ke 6. - Maize sex determination protein TASSELSEED 2. - *Sarcophaga peregrina* 25 Kd development specific protein. - *Drosophila* fat body protein P6. - A *Listeria monocytogenes* hypothetical protein encoded in the *internalins* gene region. - *Escherichia coli*

hypothetical protein yciK. - Escherichia coli hypothetical protein ydfG. - Escherichia coli
 hypothetical protein yjgI. - Escherichia coli hypothetical protein yjgU. - Escherichia coli
 hypothetical protein yohF. - Bacillus subtilis hypothetical protein yoxD. - Bacillus subtilis
 hypothetical protein ywfD. - Bacillus subtilis hypothetical protein ywfH. - Yeast hypothetical
 5 protein YIL124w. - Yeast hypothetical protein YIR035c. - Yeast hypothetical protein
 YIR036c. - Yeast hypothetical protein YKL055c. - Fission yeast hypothetical protein
 SpAC23D3.11. One of the best conserved regions which includes two perfectly conserved
 residues, a tyrosine and a lysine has been selected as a signature pattern for this family of
 proteins. The tyrosine residue participates in the catalytic mechanism.

10

Consensus pattern: [LIVSPADNK]-x(12)-Y-[PSTAGNCV]-[STAGNQCIVM]-[STAGC]-K-
 {PC}-[SAGFYR]-[LIVMSTAGD]-x(2)-[LIVMFYW]-x(3)- [LIVMFYWGAPTHQ]-
 [GSACQRHM] [Y is an active site residue] -

15

[1] Joernvall H., Persson B., Krook M., Atrian S., Gonzalez-Duarte R., Jeffery J., Ghosh D.
 Biochemistry 34:6003-6013(1995).

[2] Villarroya A., Juan E., Egestad B., Joernvall H. Eur. J. Biochem. 180:191-197(1989).

[3] Persson B., Krook M., Joernvall H. Eur. J. Biochem. 200:537-543(1991).

[4] Neidle E.L., Hartnett C., Ornston N.L., Bairoch A., Rekik M., Harayama S. Eur. J.
 20 Biochem. 204:113-120(1992).

45. (adh_short_C2) Short-chain dehydrogenases/reductases family signature

The short-chain dehydrogenases/reductases family (SDR) [1] is a very large family of
 25 enzymes, most of which are known to be NAD- or NADP-dependent oxidoreductases. As the
 first member of this family to be characterized was Drosophila alcohol dehydrogenase, this
 family used to be called [2,3,4]'insect-type', or 'short-chain' alcohol dehydrogenases. Most
 member of this family are proteins of about 250 to 300 amino acid residues. The proteins
 currently known to belong to this family are listed below. - Alcohol dehydrogenase (EC
 30 1.1.1.1) from insects such as Drosophila. - Acetoin dehydrogenase (EC 1.1.1.5) from
 Klebsiella terrigena (gene budC). - D-beta-hydroxybutyrate dehydrogenase (BDH) (EC
1.1.1.30) from mammals. - Acetoacetyl-CoA reductase (EC 1.1.1.36) from various bacterial
 species (gene phbB or phaB). - Glucose 1-dehydrogenase (EC 1.1.1.47) from Bacillus. - 3-

beta-hydroxysteroid dehydrogenase (EC 1.1.1.51) from *Comomonas testosteroni*. - 20-beta-hydroxysteroid dehydrogenase (EC 1.1.1.53) from *Streptomyces hydrogenans*. - Ribitol dehydrogenase (EC 1.1.1.56) (RDH) from *Klebsiella aerogenes*. - Estradiol 17-beta-dehydrogenase (EC 1.1.1.62) from human. - Gluconate 5-dehydrogenase (EC 1.1.1.69) from
 5 Gluconobacter oxydans (gene gno). - 3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100) from *Escherichia coli* (gene fabG) and from plants. - Retinol dehydrogenase (EC 1.1.1.105) from mammals. - 2-deoxy-d-gluconate 3-dehydrogenase (EC 1.1.1.125) from *Escherichia coli* and *Erwinia chrysanthemi* (gene kduD). - Sorbitol-6-phosphate 2-dehydrogenase (EC 1.1.1.140) from *Escherichia coli* (gene gutD) and from *Klebsiella*
 10 pneumoniae (gene sorD). - 15-hydroxyprostaglandin dehydrogenase (NAD⁺) (EC 1.1.1.141) from human. - Corticosteroid 11-beta-dehydrogenase (EC 1.1.1.146) (11-DH) from mammals. - 7-alpha-hydroxysteroid dehydrogenase (EC 1.1.1.159) from *Escherichia coli* (gene hdhA), *Eubacterium* strain VPI 12708 (gene baiA) and from *Clostridium sordellii*. - NADPH-dependent carbonyl reductase (EC 1.1.1.184) from mammals. - Tropinone
 15 reductase-I (EC 1.1.1.206) and -II (EC 1.1.1.236) from plants. - N-acylmannosamine 1-dehydrogenase (EC 1.1.1.233) from *Flavobacterium* strain 141-8. - D-arabinitol 2-dehydrogenase (ribulose forming) (EC 1.1.1.250) from fungi. - Tetrahydroxynaphthalene reductase (EC 1.1.1.252) from *Magnaporthe grisea*. - Pteridine reductase 1 (EC 1.1.1.253) (gene PTR1) from *Leishmania*. - 2,5-dichloro-2,5-cyclohexadiene-1,4-diol dehydrogenase (EC 1.1.-.-) from *Pseudomonas paucimobilis*. - Cis-1,2-dihydroxy-3,4-cyclohexadiene-1-carboxylate dehydrogenase (EC 1.3.1.-) from *Acinetobacter calcoaceticus* (gene benD) and *Pseudomonas putida* (gene xylL). - Biphenyl-2,3-dihydro-2,3-diol dehydrogenase (EC 1.3.1.-) (gene bphB) from various *Pseudomonaceae*. - Cis-toluene dihydrodiol dehydrogenase (EC 1.3.1.-) from *Pseudomonas putida* (gene todD). - Cis-benzene glycol dehydrogenase (EC
 20 1.3.1.19) from *Pseudomonas putida* (gene bnzE). - 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase (EC 1.3.1.28) from *Escherichia coli* (gene entA) and *Bacillus subtilis* (gene dhbA). - Dihydropteridine reductase (EC 1.6.99.7) (HDHPR) from mammals. - Lignin degradation enzyme ligD from *Pseudomonas naucimobilis*. - Agropine synthesis reductase from *Agrobacterium* plasmids (gene mas1). - Versicolorin reductase from *Aspergillus*
 25 parasiticus (gene VER1). - Putative keto-acyl reductases from *Streptomyces* polyketide biosynthesis operons. - A trifunctional hydratase-dehydrogenase-epimerase from the peroxisomal beta-oxidation system of *Candida tropicalis*. This protein contains two tandemly repeated 'short-chain dehydrogenase-type' domain in its N-terminal extremity. - Nodulation

protein nodG from species of *Azospirillum* and *Rhizobium* which is probably involved in the modification of the nodulation Nod factor fatty acyl chain. - Nitrogen fixation protein fixR from *Bradyrhizobium japonicum*. - *Bacillus subtilis* protein dltE which is involved in the biosynthesis of D- alanyl-lipoteichoic acid. - Human follicular variant translocation protein 1 (FVT1). - Mouse adipocyte protein p27. - Mouse protein Ke 6. - Maize sex determination protein TASSELSEED 2. - *Sarcophaga peregrina* 25 Kd development specific protein. - *Drosophila* fat body protein P6. - A *Listeria monocytogenes* hypothetical protein encoded in the internalins gene region. - *Escherichia coli* hypothetical protein yciK. - *Escherichia coli* hypothetical protein ydfG. - *Escherichia coli* hypothetical protein yjgI. - *Escherichia coli* hypothetical protein yjgU. - *Escherichia coli* hypothetical protein yohF. - *Bacillus subtilis* hypothetical protein yoxD. - *Bacillus subtilis* hypothetical protein ywfD. - *Bacillus subtilis* hypothetical protein ywfH. - Yeast hypothetical protein YIL124w. - Yeast hypothetical protein YIR035c. - Yeast hypothetical protein YIR036c. - Yeast hypothetical protein YKL055c. - Fission yeast hypothetical protein SpAC23D3.11. One of the best conserved regions which includes two perfectly conserved residues, a tyrosine and a lysine has been used as a signature pattern for this family of proteins. The tyrosine residue participates in the catalytic mechanism.

Consensus pattern: [LIVSPADNK]-x(12)-Y-[PSTAGNCV]-[STAGNQCIVM]-[STAGC]-K-{PC}-[SAGFYR]-[LIVMSTAGD]-x(2)-[LIVMFYW]-x(3)-[LIVMFYWGAPTHQ]-[GSACQRHM] [Y is an active site residue]

[1] Joernvall H., Persson B., Krook M., Atrian S., Gonzalez-Duarte R., Jeffery J., Ghosh D. *Biochemistry* 34:6003-6013(1995).

[2] Villarroja A., Juan E., Egestad B., Joernvall H. *Eur. J. Biochem.* 180:191-197(1989).

[3] Persson B., Krook M., Joernvall H. *Eur. J. Biochem.* 200:537-543(1991).

[4] Neidle E.L., Hartnett C., Ornston N.L., Bairoch A., Rekik M., Harayama S. *Eur. J. Biochem.* 204:113-120(1992).

46. (adh_zinc) Zinc-containing alcohol dehydrogenases signatures

Alcohol dehydrogenase (EC 1.1.1.1) (ADH) catalyzes the reversible oxidation of ethanol to acetaldehyde with the concomitant reduction of NAD [1]. Currently three, structurally and

catalytically, different types of alcohol dehydrogenases are known: - Zinc-containing 'long-chain' alcohol dehydrogenases. - Insect-type, or 'short-chain' alcohol dehydrogenases. - Iron-containing alcohol dehydrogenases. Zinc-containing ADH's [2,3] are dimeric or tetrameric enzymes that bind two atoms of zinc per subunit. One of the zinc atom is essential for catalytic activity while the other is not. Both zinc atoms are coordinated by either cysteine or histidine residues; the catalytic zinc is coordinated by two cysteines and one histidine. Zinc-containing ADH's are found in bacteria, mammals, plants, and in fungi. In most species there are more than one isozyme (for example, human have at least six isozymes, yeast have three, etc.). A number of other zinc-dependent dehydrogenases are closely related to zinc ADH [4], these are: - Xylitol dehydrogenase (EC 1.1.1.9) (D-xylulose reductase). - Sorbitol dehydrogenase (EC 1.1.1.14). - Aryl-alcohol dehydrogenase (EC 1.1.1.90) (benzyl alcohol dehydrogenase). - Threonine 3-dehydrogenase (EC 1.1.1.103). - Cinnamyl-alcohol dehydrogenase (EC 1.1.1.195) (CAD) [5]. CAD is a plant enzyme involved in the biosynthesis of lignin. - Galactitol-1-phosphate dehydrogenase (EC 1.1.1.251). - Pseudomonas putida 5-exo-alcohol dehydrogenase (EC 1.1.1.-) [6]. - Escherichia coli starvation sensing protein *rspB*. - Escherichia coli hypothetical protein *yjgB*. - Escherichia coli hypothetical protein *yjgV*. - Escherichia coli hypothetical protein *yjiN*. - Yeast hypothetical protein YAL060w (FUN49). - Yeast hypothetical protein YAL061w (FUN50). - Yeast hypothetical protein YCR105w. The pattern that has been developed to detect this class of enzymes is based on a conserved region that includes a histidine residue which is the second ligand of the catalytic zinc atom. This family also includes NADP-dependent quinone oxidoreductase (EC 1.6.5.5), an enzyme found in bacteria (gene *qor*), in yeast and in mammals where, in some species such as rodents, it has been recruited as an eye lens protein and is known as zeta-crystallin [7]. The sequence of quinone oxidoreductase is distantly related to that other zinc-containing alcohol dehydrogenases and it lacks the zinc-ligand residues. The torpedo fish and mammalian synaptic vesicle membrane protein *vat-1* is related to *qor*. A specific pattern has been developed for this subfamily.

Consensus pattern: G-H-E-x(2)-G-x(5)-[GA]-x(2)-[IVSAC] [H is a zinc ligand]

Consensus pattern: [GSD]-[DEQH]-x(2)-L-x(3)-[SA](2)-G-G-x-G-x(4)-Q-x(2)-[KR]-

[1] Branden C.-I., Joernvall H., Eklund H., Furugren B. (In) The Enzymes (3rd edition) 11:104-190(1975).

- [2] Joernvall H., Persson B., Jeffery J. *Eur. J. Biochem.* 167:195-201(1987).
- [3] Sun H.-W., Plapp B.V. *J. Mol. Evol.* 34:522-535(1992).
- [4] Persson B., Hallborn J., Walfridsson M., Hahn-Haegerdal B., Keraenen S., Penttilae M., Joernvall H. *FEBS Lett.* 324:9-14(1993).
- 5 [5] Knight M.E., Halpin C., Schuch W. *Plant Mol. Biol.* 19:793-801(1992).
- [6] Koga H., Aramaki H., Yamaguchi E., Takeuchi K., Horiuchi T., Gunsalus I.C. *J. Bacteriol.* 166:1089-1095(1986).
- [7] Joernvall H., Persson B., Du Bois G., Lavers G.C., Chen J.H., Gonzalez P., Rao P.V., Zigler J.S. Jr. *FEBS Lett.* 322:240-244(1993).

10

47. (aldehyd) Aldehyde dehydrogenases active sites

Aldehyde dehydrogenases (EC 1.2.1.3 and EC 1.2.1.5) are enzymes which oxidize a wide variety of aliphatic and aromatic aldehydes. In mammals at least four different forms of the enzyme are known [1]: class-1 (or Ald C) a tetrameric cytosolic enzyme, class-2 (or Ald M) a tetrameric mitochondrial enzyme, class-3 (or Ald D) a dimeric cytosolic enzyme, and class IV a microsomal enzyme. Aldehyde dehydrogenases have also been sequenced from fungal and bacterial species. A number of enzymes are known to be evolutionary related to aldehyde dehydrogenases; these enzymes are listed below. - Plants and bacterial betaine-aldehyde dehydrogenase (EC 1.2.1.8) [2], an enzyme that catalyzes the last step in the biosynthesis of betaine. - Plants and bacterial NADP-dependent glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.9). - *Escherichia coli* succinate-semialdehyde dehydrogenase (NADP+) (EC 1.2.1.16) (gene *gabD*) [3], which reduces succinate semialdehyde into succinate. - *Escherichia coli* lactaldehyde dehydrogenase (EC 1.2.1.22) (gene *ald*) [4]. - Mammalian succinate semialdehyde dehydrogenase (NAD+) (EC 1.2.1.24). - *Escherichia coli* phenylacetaldehyde dehydrogenase (EC 1.2.1.39). - *Escherichia coli* 5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase (gene *hpcC*). - *Pseudomonas putida* 2-hydroxymuconic semialdehyde dehydrogenase [5] (genes *dmpC* and *xylG*), an enzyme in the meta-cleavage pathway for the degradation of phenols, cresols and catechol. - Bacterial and mammalian methylmalonate-semialdehyde dehydrogenase (MMSDH) (EC 1.2.1.27) [6], an enzyme involved in the distal pathway of valine catabolism. - Yeast delta-1-pyrroline-5-carboxylate dehydrogenase (EC 1.5.1.12) [7] (gene *PUT2*), which converts proline to glutamate. - Bacterial multifunctional putA protein, which contains a delta-1-pyrroline- 5-

30

carboxylate dehydrogenase domain. - 26G, a garden pea protein of unknown function which is induced by dehydration of shoots [8]. - Mammalian formyltetrahydrofolate dehydrogenase (EC 1.5.1.6) [9]. This is a cytosolic enzyme responsible for the NADP-dependent decarboxylative reduction of 10-formyltetrahydrofolate into tetrahydrofolate. It is an protein of about 900 amino acids which consist of three domains; the C- terminal domain (480 residues) is structurally and functionally related to aldehyde dehydrogenases. - Yeast hypothetical protein YBR006w. - Yeast hypothetical protein YER073w. - Yeast hypothetical protein YHR039c. - *Caenorhabditis elegans* hypothetical protein F01F1.6.A glutamic acid and a cysteine residue have been implicated in the catalytic activity of mammalian aldehyde dehydrogenase. These residues are conserved in all the enzymes of this family. Two patterns have been derived for this family, one for each of the active site residues.

Consensus pattern: [LIVMFGA]-E-[LIMSTAC]-[GS]-G-[KNLM]-[SADN]-[TAPFV] [E is the active site residue]-

Consensus pattern: [FYLV A]-x(3)-G-[QE]-x-C-[LIVMGSTANC]-[AGCN]-x-[GSTADNEKR] [C is the active site residue]

[1] Hempel J., Harper K., Lindahl R. *Biochemistry* 28:1160-1167(1989).

[2] Weretilnyk E.A., Hanson A.D. *Proc. Natl. Acad. Sci. U.S.A.* 87:2745-2749(1990).

[3] Niegemann E., Schulz A., Bartsch K. *Arch. Microbiol.* 160:454-460(1993).

[4] Hidalgo E., Chen Y.-M., Lin E.C.C., Aguilar J. J. *Bacteriol.* 173:6118-6123(1991).

[5] Nordlund I., Shingler V. *Biochim. Biophys. Acta* 1049:227-230(1990).

[6] Steele M.I., Lorenz D., Hatter K., Park A., Sokatch J.R. *J. Biol. Chem.* 267:13585-13592(1992).

[7] Krzywicki K.A., Brandriss M.C. *Mol. Cell. Biol.* 4:2837-2842(1984).

[8] Guerrero F.D., Jones J.T., Mullet J.E. *Plant Mol. Biol.* 15:11-26(1990).

[9] Cook R.J., Lloyd R.S., Wagner C. *J. Biol. Chem.* 266:4965-4973(1991).

48. Aldo/keto reductase family signatures

The aldo-keto reductase family [1,2] groups together a number of structurally and functionally related NADPH-dependent oxidoreductases as well as some other proteins. The proteins known to belong to this family are: - Aldehyde reductase (EC 1.1.1.2). - Aldose

reductase (EC 1.1.1.21). - 3-alpha-hydroxysteroid dehydrogenase (EC 1.1.1.50), which terminates androgen action by converting 5-alpha-dihydrotestosterone to 3-alpha-androstanediol. - Prostaglandin F synthase (EC 1.1.1.188) which catalyzes the reduction of prostaglandins H2 and D2 to F2-alpha. - D-sorbitol-6-phosphate dehydrogenase (EC 1.1.1.200) from apple. - Morphine 6-dehydrogenase (EC 1.1.1.218) from *Pseudomonas putida* plasmid pMDH7.2 (gene morA). - Chlordecone reductase (EC 1.1.1.225) which reduces the pesticide chlordecone (kepone) to the corresponding alcohol. - 2,5-diketo-D-gluconic acid reductase (EC 1.1.1.-) which catalyzes the reduction of 2,5-diketogluconic acid to 2-keto-L-gulonic acid, a key intermediate in the production of ascorbic acid. - NAD(P)H-dependent xylose reductase (EC 1.1.1.-) from the yeast *Pichia stipitis*. This enzyme reduces xylose into xylitol. - Trans-1,2-dihydrobenzene-1,2-diol dehydrogenase (EC 1.3.1.20). - 3-oxo-5-beta-steroid 4-dehydrogenase (EC 1.3.99.6) which catalyzes the reduction of delta(4)-3-oxosteroids. - A soybean reductase, which co-acts with chalcone synthase in the formation of 4,2',4'-trihydroxychalcone. - Frog eye lens rho crystallin. - Yeast GCY protein, whose function is not known. - *Leishmania major* P110/11E protein. P110/11E is a developmentally regulated protein whose abundance is markedly elevated in promastigotes compared with amastigotes. Its exact function is not yet known. - *Escherichia coli* hypothetical protein yafB. - *Escherichia coli* hypothetical protein yghE. - Yeast hypothetical protein YBR149w. - Yeast hypothetical protein YHR104w. - Yeast hypothetical protein YJR096w. These proteins have all about 300 amino acid residues. Three consensus patterns have been developed that are specific to this family of proteins. The first one is located in the N-terminal section of these proteins. The second pattern is located in the central section. The third pattern, located in the C-terminal, is centered on a lysine residue whose chemical modification, in aldose and aldehydereductases, affect the catalytic efficiency.

Consensus pattern: G-[FY]-R-[HSAL]-[LIVMF]-D-[STAGC]-[AS]-x(5)-E-x(2)-[LIVM]-G -
 Consensus pattern: [LIVMFY]-x(9)-[KREQ]-x-[LIVM]-G-[LIVM]-[SC]-N-[FY]-
 Consensus pattern: [LIVM]-[PAIV]-[KR]-[ST]-x(4)-R-x(2)-[GSTAEQK]-[NSL]-x(2)-
 [LIVMFA] [K is a putative active site residue]-

[1] Bohren K.M., Bullock B., Wermuth B., Gabbay K.H. J. Biol. Chem. 264:9547-9551(1989).

[2] Bruce N.C., Willey D.L., Coulson A.F.W., Jeffery J. Biochem. J. 299:805-811(1994).

49. Alpha amylase. This family is classified as family 13 of the glycosyl hydrolases. The structure is an 8 stranded alpha/beta barrel, interrupted by a ~70 a.a. calcium-binding domain protruding between beta strand 3 and alpha helix 3, and a carboxyl-terminal Greek key beta-barrel domain.

[1] Larson SB, Greenwood A, Cascio D, Day J, McPherson A, J Mol Biol 1994;235:1560-1584.

50. Aminotransferases class-I pyridoxal-phosphate attachment site

Aminotransferases share certain mechanistic features with other pyridoxal- phosphate dependent enzymes, such as the covalent binding of the pyridoxal- phosphate group to a lysine residue. On the basis of sequence similarity, these various enzymes can be grouped [1,2] into subfamilies. One of these, called class-I, currently consists of the following enzymes: - Aspartate aminotransferase (AAT) (EC 2.6.1.1). AAT catalyzes the reversible transfer of the amino group from L-aspartate to 2-oxoglutarate to form oxaloacetate and L-glutamate. In eukaryotes, there are two AAT isozymes: one is located in the mitochondrial matrix, the second is cytoplasmic. In prokaryotes, only one form of AAT is found (gene aspC). - Tyrosine aminotransferase (EC 2.6.1.5) which catalyzes the first step in tyrosine catabolism by reversibly transferring its amino group to 2- oxoglutarate to form 4-hydroxyphenylpyruvate and L-glutamate. - Aromatic aminotransferase (EC 2.6.1.57) involved in the synthesis of Phe, Tyr, Asp and Leu (gene tyrB). - 1-aminocyclopropane-1-carboxylate synthase (EC 4.4.1.14) (ACC synthase) from plants. ACC synthase catalyzes the first step in ethylene biosynthesis. - Pseudomonas denitrificans cobC, which is involved in cobalamin biosynthesis. - Yeast hypothetical protein YJL060w. The sequence around the pyridoxal-phosphate attachment site of this class of enzyme is sufficiently conserved to allow the creation of a specific pattern.

Consensus pattern: [GS]-[LIVMFYTAC]-[GSTA]-K-x(2)-[GSALVN]-[LIVMFA]-x-[GNAR]- x-R-[LIVMA]-[GA] [K is the pyridoxal-P attachment site]

[1] Bairoch A. Unpublished observations (1992).

[2] Sung M.H., Tanizawa K., Tanaka H., Kuramitsu S., Kagamiyama H., Hirotsu K., Okamoto A., Higuchi T., Soda K. J. Biol. Chem. 266:2567-2572(1991).

5

51. Aminotransferases class-II pyridoxal-phosphate attachment site

Aminotransferases share certain mechanistic features with other pyridoxal- phosphate dependent enzymes, such as the covalent binding of the pyridoxal- phosphate group to a lysine residue. On the basis of sequence similarity, these various enzymes can be grouped [1] into subfamilies. One of these, called class-II, currently consists of the following enzymes: - Glycine acetyltransferase (EC 2.3.1.29), which catalyzes the addition of acetyl-CoA to glycine to form 2-amino-3-oxobutanoate (gene kbl). - 5-aminolevulinic acid synthase (EC 2.3.1.37) (delta-ALA synthase), which catalyzes the first step in heme biosynthesis via the Shemin (or C4) pathway, i.e. the addition of succinyl-CoA to glycine to form 5-aminolevulinate. - 8-amino-7-oxononanoate synthase (EC 2.3.1.47) (7-KAP synthetase), a bacterial enzyme (gene bioF) which catalyzes an intermediate step in the biosynthesis of biotin: the addition of 6-carboxy-hexanoyl-CoA to alanine to form 8-amino-7-oxononanoate. - Histidinol-phosphate aminotransferase (EC 2.6.1.9), which catalyzes the eighth step in histidine biosynthetic pathway: the transfer of an amino group from 3-(imidazol-4-yl)-2-oxopropyl phosphate to glutamic acid to form histidinol phosphate and 2-oxoglutarate. - Serine palmitoyltransferase (EC 2.3.1.50) from yeast (genes LCB1 and LCB2), which catalyzes the condensation of palmitoyl-CoA and serine to form 3- ketosphinganine. The sequence around the pyridoxal-phosphate attachment site of this class of enzyme is sufficiently conserved to allow the creation of a specific pattern

25

Consensus pattern: T-[LIVMFYW]-[STAG]-K-[SAG]-[LIVMFYWR]-[SAG]-x(2)-[SAG]
[K is the pyridoxal-P attachment site]-

[1] Bairoch A. Unpublished observations (1991).

30

52. Aminotransferases class-III pyridoxal-phosphate attachment site

Aminotransferases share certain mechanistic features with other pyridoxal- phosphate dependent enzymes, such as the covalent binding of the pyridoxal- phosphate group to a lysine residue. On the basis of sequence similarity, these various enzymes can be grouped [1,2] into subfamilies. One of these, called class-III, currently consists of the following

5 enzymes: - Acetylornithine aminotransferase (EC 2.6.1.11) which catalyzes the transfer of an amino group from acetylornithine to alpha-ketoglutarate, yielding N-acetyl-glutamic-5-semi-aldehyde and glutamic acid. - Ornithine aminotransferase (EC 2.6.1.13), which catalyzes the transfer of an amino group from ornithine to alpha-ketoglutarate, yielding glutamic-5- semi-aldehyde and glutamic acid. - Omega-amino acid--pyruvate aminotransferase (EC 2.6.1.18),

10 which catalyzes transamination between a variety of omega-amino acids, mono- and diamines, and pyruvate. It plays a pivotal role in omega amino acids metabolism. - 4-aminobutyrate aminotransferase (EC 2.6.1.19) (GABA transaminase), which catalyzes the transfer of an amino group from GABA to alpha-ketoglutarate, yielding succinate semialdehyde and glutamic acid. - DAPA aminotransferase (EC 2.6.1.62), a bacterial enzyme (gene bioA) which catalyzes an intermediate step in the biosynthesis of biotin, the

15 transamination of 7-keto-8-aminopelargonic acid (7-KAP) to form 7,8- diaminopelargonic acid (DAPA). - 2,2-dialkylglycine decarboxylase (EC 4.1.1.64), a *Pseudomonas cepacia* enzyme (gene dgdA) that catalyzes the decarboxylating amino transfer of 2,2-dialkylglycine and pyruvate to dialkyl ketone, alanine and carbon dioxide. - Glutamate-1-semialdehyde

20 aminotransferase (EC 5.4.3.8) (GSA). GSA is the enzyme involved in the second step of porphyrin biosynthesis, via the C5 pathway. It transfers the amino group on carbon 2 of glutamate-1- semialdehyde to the neighbouring carbon, to give delta-aminolevulinic acid. - *Bacillus subtilis* aminotransferase yhxA. - *Bacillus subtilis* aminotransferase yodT. - *Haemophilus influenzae* aminotransferase HI0949. - *Caenorhabditis elegans* aminotransferase

25 T01B11.2. The sequence around the pyridoxal-phosphate attachment site of this class of enzyme is sufficiently conserved to allow the creation of a specific pattern.

Consensus pattern: [LIVMFYWC](2)-x-D-E-[IVA]-x(2)-G-[LIVMFAGC]-x(0,1)-
 [RSACLI]-x-[GSAD]-x(12,16)-D-[LIVMFC]-[LIVMFYSTA]-x(2)- [GSA]-K-x(3)-
 30 [GSTADNV]-[GSAC] [K is the pyridoxal-P attachment site]-

[1] Bairoch A. Unpublished observations (1992).[2] Yonaha K., Nishie M., Aibara S. J. Biol. Chem. 267:12506-12510(1992).

53. Ank repeat. There's no clear separation between noise and signal on the HMM search
Ankyrin repeats generally consist of a beta, alpha, alpha, beta order of secondary structures.

5 The repeats associate to form a higher order structure.

[1] A, Holak TA, FEBS Lett 1997;401:127-132.

[2] Lux SE, John KM, Bennett V, Nature 1990;345:736-739.

10

54. Aminotransferases class-IV signature

Aminotransferases share certain mechanistic features with other pyridoxal-phosphate
dependent enzymes, such as the covalent binding of the pyridoxal-phosphate group to a
lysine residue. On the basis of sequence similarity, these various enzymes can be grouped
[1,2] into subfamilies. One of these, called class-IV, currently consists of the following
enzymes:

- Branched-chain amino-acid aminotransferase (EC 2.6.1.42) (transaminase B), a
bacterial (gene *ilvE*) and eukaryotic enzyme which catalyzes the reversible
transfer of an amino group from 4-methyl-2-oxopentanoate to glutamate, to form
leucine and 2-oxoglutarate.
- D-alanine aminotransferase (EC 2.6.1.21). A bacterial enzyme which catalyzes the
transfer of the amino group from D-alanine (and other D-amino acids) to 2-
oxoglutarate, to form pyruvate and D-aspartate.
- 4-amino-4-deoxychorismate (ADC) lyase (gene *pabC*). A bacterial enzyme that
converts ADC into 4-aminobenzoate (PABA) and pyruvate.

The above enzymes are proteins of about 270 to 415 amino-acid residues that share a
few regions of sequence similarity. Surprisingly, the best-conserved region does not include
the lysine residue to which the pyridoxal-phosphate group is known to be attached, in *ilvE*.
The region that has been selected as a signature pattern is located some 40 residues at the C-
terminus side of the PIP-lysine

Consensus pattern: E-x-[STAGCI]-x(2)-N-[LIVMFAC]-[FY]-x(6,12)-[LIVMF]-x-T-x(6,8)-
[LIVM]-x-[GS]-[LIVM]-x-[KR]-

[1] Green J.M., Merkel W.K., Nichols B.P. J. Bacteriol. 174:5317-5323(1992).

[2] Bairoch A. Unpublished observations (1992).

55. Aminotransferases class-V pyridoxal-phosphate attachment site

Aminotransferases share certain mechanistic features with other pyridoxal- phosphate dependent enzymes, such as the covalent binding of the pyridoxal- phosphate group to a lysine residue. On the basis of sequence similarity, these various enzymes can be grouped [1,2] into subfamilies. One of these, called class-V, currently consists of the following enzymes: - Phosphoserine aminotransferase (EC 2.6.1.52), an enzyme which catalyzes the reversible interconversion of phosphoserine and 2-oxoglutarate to 3-phosphonooxypyruvate and glutamate. It is required both in the major phosphorylated pathway of serine biosynthesis and in pyridoxine biosynthesis. The bacterial enzyme (gene serC) is highly similar to a rabbit endometrial progesterone-induced protein (EPIP), which is probably a phosphoserine aminotransferase [3]. - Serine--glyoxylate aminotransferase (EC 2.6.1.45) (SGAT) (gene sgaA) from *Methylobacterium extorquens*. - Serine--pyruvate aminotransferase (EC 2.6.1.51). This enzyme also acts as an alanine--glyoxylate aminotransferase (EC 2.6.1.44). In vertebrates, it is located in the peroxisomes and/or mitochondria. - Isopenicillin N epimerase (gene cefD). This enzyme is involved in the biosynthesis of cephalosporin antibiotics and catalyzes the reversible isomerization of isopenicillin N and penicillin N. - NifS, a protein of the nitrogen fixation operon of some bacteria and cyanobacteria. The exact function of nifS is not yet known. A highly similar protein has been found in fungi (gene NFS1 or SPL1). - The small subunit of cyanobacterial soluble hydrogenase (EC 1.12.-.-). - Hypothetical protein ycbU from *Bacillus subtilis*. - Hypothetical protein YFL030w from yeast. The sequence around the pyridoxal-phosphate attachment site of this class of enzyme is sufficiently conserved to allow the creation of a specific pattern.

Consensus pattern: [LIVFYCHT]-[DGH]-[LIVMFYAC]-[LIVMFYA]-x(2)-[GSTAC]-[GSTA]- [HQR]-K-x(4,6)-G-x-[GSAT]-x-[LIVMFYSAC] [K is the pyridoxal-P attachment site]-

[1] Ouzounis C., Sander C. FEBS Lett. 322:159-164(1993).

[2] Bairoch A. Unpublished observations (1992).

[3] van der Zel A., Lam H.-M., Winkler M.E. *Nucleic Acids Res.* 17:8379-8379(1989).

56. Annexins repeated domain signature

5 Annexins [1 to 6] are a group of calcium-binding proteins that associate reversibly with membranes. They bind to phospholipid bilayers in the presence of micromolar free calcium concentration. The binding is specific for calcium and for acidic phospholipids. Annexins have been claimed to be involved in cytoskeletal interactions, phospholipase inhibition, intracellular signalling, anticoagulation, and membrane fusion. Each of these proteins consist
10 of an N-terminal domain of variable length followed by four or eight copies of a conserved segment of sixty one residues. The repeat (sometimes known as an 'endonexin fold') consists of five alpha-helices that are wound into a right-handed superhelix [7]. The proteins known to belong to the annexin family are listed below: - Annexin I (Lipocortin 1) (Calpactin 2) (p35) (Chromobindin 9). - Annexin II (Lipocortin 2) (Calpactin 1) (Protein I) (p36) (Chromobindin
15 8). - Annexin III (Lipocortin 3) (PAP-III). - Annexin IV (Lipocortin 4) (Endonexin I) (Protein II) (Chromobindin 4). - Annexin V (Lipocortin 5) (Endonexin 2) (VAC-alpha) (Anchorin CII) (PAP-I). - Annexin VI (Lipocortin 6) (Protein III) (Chromobindin 20) (p68) (p70). This is the only known annexin that contains 8 (instead of 4) repeats. - Annexin VII (Synexin). - Annexin VIII (Vascular anticoagulant-beta) (VAC-beta). - Annexin IX from *Drosophila*. -
20 Annexin X from *Drosophila*. - Annexin XI (Calcyclin-associated annexin) (CAP-50). - Annexin XII from *Hydra vulgaris*. - Annexin XIII (Intestine-specific annexin) (ISA). The signature pattern for this domain spans positions 9 to 61 of the repeat and includes the only perfectly conserved residue (an arginine in position 22)-

25 Consensus pattern: [TG]-[STV]-x(8)-[LIVMF]-x(2)-R-x(3)-[DEQNH]-x(7)-[IFY]- x(7)-[LIVMF]-x(3)-[LIVMF]-x(11)-[LIVMFA]-x(2)-[LIVMF]-

[1] Raynal P., Pollard H.B. *Biochim. Biophys. Acta* 1197:63-93(1994).

[2] Barton G.J., Newman R.H., Freemont P.S., Crumpton M.J. *Eur. J. Biochem.* 198:749-
30 760(1991).

[3] Burgoyne R.D., Geisow M.J. *Cell Calcium* 10:1-10(1989).

[4] Haigler H.T., Fitch J.M., Jones J.M., Schlaepfer D.D. *Trends Biochem. Sci.* 14:48-
50(1989).

- [5] Klee C.B. Biochemistry 27:6645-6653(1988).
- [6] Smith P.D., Moss S.E. Trends Genet. 10:241-246(1994).
- [7] Huber R., Roemisch J., Paques E.-P. EMBO J. 9:3867-3874(1990).
- [8] Fiedler K., Simons K. Trends Biochem. Sci. 20:177-178(1995).

5

57. (arf_1) ADP-ribosylation factors family signature

ADP-ribosylation factors (ARF) [1,2,3,4] are 20 Kd GTP-binding proteins involved in protein trafficking. They may modulate vesicle budding and uncoating within the Golgi apparatus. ARF's also act as allosteric activators of cholera toxin ADP-ribosyltransferase activity. They are evolutionary conserved and present in all eukaryotes. At least six forms of ARF are present in mammals and three in budding yeast. The ARF family also includes proteins highly related to ARF's but which lack the cholera toxin cofactor activity, they are collectively known as ARL's (ARF-like). ARD1 is a 64 Kd mammalian protein of unknown biological function that contains an ARF domain at its C-terminal extremity. Proteins from the ARF family are generally included in the RAS 'superfamily' of small GTP-binding proteins [5], but they are only slightly related to the other RAS proteins. They also differ from RAS proteins in that they lack cysteine residues at their C-termini and are therefore not subject to prenylation. The ARFs are N-terminally myristoylated (the ARLs have not yet been shown to be modified in such a fashion). A conserved region in the C-terminal part of ARF's and ARL's has been selected as a signature pattern.

Consensus pattern: [HRQT]-x-[FYWI]-x-[LIVM]-x(4)-A-x(2)-G-x(2)-[LIVM]-x(2)- [GSA]-[LIVMF]-x-[WK]-[LIVM]-

Note: proteins belonging to this family also contain a copy of the ATP/GTP- binding motif 'A' (P-loop) (see <[PDOC00017](#)

- [1] Boman A.L., Kahn R.A. Trends Biochem. Sci. 20:147-150(1995).
- [2] Moss J., Vaughan M. Cell. Signal. 4:367-399(1993).
- [3] Moss J., Vaughan M. Prog. Nucleic Acid Res. Mol. Biol. 45:47-65(1993).
- [4] Amor J.C., Harrison D.H., Kahn R.A., Ringe D. Nature 372:704-708(1994).
- [5] Valencia A., Chardin P., Wittinghofer A., Sander C. Biochemistry 30:4637-4648(1991).

(arf_2) ATP/GTP-binding site motif A (P-loop)

From sequence comparisons and crystallographic data analysis it has been shown

[1,2,3,4,5,6] that an appreciable proportion of proteins that bind ATP or GTP share a number of more or less conserved sequence motifs. The best conserved of these motifs is a glycine-rich region, which typically forms a flexible loop between a beta-strand and an alpha-helix.

This loop interacts with one of the phosphate groups of the nucleotide. This sequence motif is generally referred to as the 'A' consensus sequence [1] or the 'P-loop' [5]. There are numerous ATP- or GTP-binding proteins in which the P-loop is found. A number of protein families for which the relevance of the presence of such motif has been noted are listed below: - ATP

synthase alpha and beta subunits (see <PDOC00137>). - Myosin heavy chains. - Kinesin heavy chains and kinesin-like proteins (see <PDOC00343>). - Dynamins and dynamin-like proteins (see <PDOC00362>). - Guanylate kinase (see <PDOC00670>). - Thymidine kinase (see <PDOC00524>). - Thymidylate kinase (see <PDOC01034>). - Shikimate kinase (see <PDOC00868>). - Nitrogenase iron protein family (nifH/frxC) (see <PDOC00580>). - ATP-binding proteins involved in 'active transport' (ABC transporters) [7] (see <PDOC00185>). - DNA and RNA helicases [8,9,10]. - GTP-binding elongation factors (EF-Tu, EF-1alpha, EF-G, EF-2, etc.). - Ras family of GTP-binding proteins (Ras, Rho, Rab, Ral, Ypt1, SEC4, etc.). - Nuclear protein ran (see <PDOC00859>). - ADP-ribosylation factors family (see <PDOC00781>). - Bacterial dnaA protein (see <PDOC00771>). - Bacterial recA protein (see <PDOC00131>). - Bacterial recF protein (see <PDOC00539>). - Guanine nucleotide-binding proteins alpha subunits (Gi, Gs, Gt, G0, etc.). - DNA mismatch repair proteins mutS family (See <PDOC00388>). - Bacterial type II secretion system protein E (see <PDOC00567>). Not

all ATP- or GTP-binding proteins are picked-up by this motif. A number of proteins escape detection because the structure of their ATP-binding site is completely different from that of the P-loop. Examples of such proteins are the E1-E2 ATPases or the glycolytic kinases. In other ATP- or GTP-binding proteins the flexible loop exists in a slightly different form; this is the case for tubulins or protein kinases. A special mention must be reserved for adenylate kinase, in which there is a single deviation from the P-loop pattern: in the last position Gly is found instead of Ser or Thr.

Consensus pattern: [AG]-x(4)-G-K-[ST]-

[1] Walker J.E., Saraste M., Runswick M.J., Gay N.J. EMBO J. 1:945-951(1982).

- [2] Moller W., Amons R. FEBS Lett. 186:1-7(1985).
 [3] Fry D.C., Kuby S.A., Mildvan A.S. Proc. Natl. Acad. Sci. U.S.A. 83:907-911(1986).
 [4] Dever T.E., Glynnias M.J., Merrick W.C. Proc. Natl. Acad. Sci. U.S.A. 84:1814-1818(1987).
 5 [5] Saraste M., Sibbald P.R., Wittinghofer A. Trends Biochem. Sci. 15:430-434(1990).
 [6] Koonin E.V. J. Mol. Biol. 229:1165-1174(1993).
 [7] Higgins C.F., Hyde S.C., Mimmack M.M., Gileadi U., Gill D.R., Gallagher M.P. J. Bioenerg. Biomembr. 22:571-592(1990).
 [8] Hodgman T.C. Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).
 10 [9] Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K., Schnier J., Slonimski P.P. Nature 337:121-122(1989).
 [10] Gorbalenya A.E., Koonin E.V., Donchenko A.P., Blinov V.M. Nucleic Acids Res. 17:4713-4730(1989).

58. Arginase family signatures

The following enzymes have been shown [1] to be evolutionary related: - Arginase (EC 3.5.3.1), a ubiquitous enzyme which catalyzes the degradation of arginine to ornithine and urea [2]. - Agmatinase (EC 3.5.3.11) (agmatine ureohydrolase), a prokaryotic enzyme (gene speB) that catalyzes the hydrolysis of agmatine into putrescine and urea. - Formiminoglutamase (EC 3.5.3.8) (formiminoglutamate hydrolase), a prokaryotic enzyme (gene hutG) that hydrolyzes N-formimino-glutamate into glutamate and formamide. - Hypothetical proteins from methanogenic archaebacteria. These enzymes are proteins of about 300 amino-acid residues. Three conserved regions that contain charged residues which are involved in the binding of the two manganese ions [3] can be used as signature patterns.-

Consensus pattern: [LIVMF]-G-G-x-H-x-[LIVMT]-[STAV]-x-[PAG]-x(3)-[GSTA] [H binds manganese]-

Consensus pattern: [LIVM](2)-x-[LIVMFY]-D-[AS]-H-x-D [The two D's and the H bind manganese]-

Consensus pattern: [ST]-[LIVMFY]-D-[LIVM]-D-x(3)-[PAQ]-x(3)-P-[GSA]-x(7)-G [The two D's bind manganese]

- [1] Ouzounis C., Kyripides N.C. J. Mol. Evol. 39:101-104(1994).
 [2] Jenkinson C.P., Grody W.W., Cederbaum S.D. Comp. Biochem. Physiol. 114B:107-132(196).
 [3] Kanyo Z.F., Scolnick L.R., Ash D.E., Christianson D.W. Nature 383:554-557(1996).

5

59. (asp) Eukaryotic and viral aspartyl proteases active site

Aspartyl proteases, also known as acid proteases, (EC 3.4.23.-) are a widely distributed family of proteolytic enzymes [1,2,3] known to exist invertebrates, fungi, plants, retroviruses and some plant viruses. Aspartate proteases of eukaryotes are monomeric enzymes which consist of two domains. Each domain contains an active site centered on a catalytic aspartyl residue. The two domains most probably evolved from the duplication of an ancestral gene encoding a primordial domain. Currently known eukaryotic aspartyl proteases are: - Vertebrate gastric pepsins A and C (also known as gastricsin). - Vertebrate chymosin (rennin), involved in digestion and used for making cheese. - Vertebrate lysosomal cathepsins D (EC 3.4.23.5) and E (EC 3.4.23.34). - Mammalian renin (EC 3.4.23.15) whose function is to generate angiotensin I from angiotensinogen in the plasma. - Fungal proteases such as aspergillopepsin A (EC 3.4.23.18), candidapepsin (EC 3.4.23.24), mucoropepsin (EC 3.4.23.23) (mucor rennin), endothiapepsin (EC 3.4.23.22), polyporopepsin (EC 3.4.23.29), and rhizopuspepsin (EC 3.4.23.21). - Yeast saccharopepsin (EC 3.4.23.25) (proteinase A) (gene PEP4). PEP4 is implicated in posttranslational regulation of vacuolar hydrolases. - Yeast barrier pepsin (EC 3.4.23.35) (gene BAR1); a protease that cleaves alpha-factor and thus acts as an antagonist of the mating pheromone. - Fission yeast sxa1 which is involved in degrading or processing the mating pheromones. Most retroviruses and some plant viruses, such as badnaviruses, encode for anaspartyl protease which is an homodimer of a chain of about 95 to 125 amino acids. In most retroviruses, the protease is encoded as a segment of apolypolyprotein which is cleaved during the maturation process of the virus. It is generally part of the pol polyprotein and, more rarely, of the gagpolyprotein. Conservation of the sequence around the two aspartates of eukaryotic aspartyl proteases and around the single active site of the viral proteases allows us to develop a single signature pattern for both groups of protease.

Consensus pattern: [LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-[STAPDENQ]- x-[LIVMFSTNC]-x-[LIVMFGTA] [D is the active site residue]

Note: these proteins belong to families A1 and A2 in the classification of peptidases [4,E1

[1] Foltmann B. Essays Biochem. 17:52-84(1981).

[2] Davies D.R. Annu. Rev. Biophys. Chem. 19:189-215(1990).

5 [3] Rao J.K.M., Erickson J.W., Wlodawer A. Biochemistry 30:4663-4671(1991).

[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:105-120(1995).

60. (BIRA) Biotin repressor

10 [1] Wilson KP, Shewchuk LM, Brennan RG, Otsuka AJ, Matthews BW; Proc Natl Acad Sci USA 1992;89:9257-9261.

61. BTB/POZ domain

15 The BTB (for BR-C, ttk and bab) [1] or POZ (for Pox virus and Zinc finger)[2] domain is present near the N-terminus of a fraction of zinc finger

(zf-C2H2) proteins and in proteins that contain the Kelch motif

such as Kelch and a family of pox virus proteins. The BTB/POZ domain mediates homomeric dimerisation and in some instances heteromeric dimerisation [2].The structure of the dimerised PLZF BTB/POZ domain has been solved and consists of a tightly intertwined homodimer. The central scaffolding of the protein is made up of a cluster of alpha-helices flanked by short beta-sheets at both the top and bottom of the molecule [3]. POZ domains from several zinc finger proteins have been shown to mediate transcriptional repression and to interact with components of histone deacetylase co-repressor complexes including N-CoR and SMRT [4,5,6]. The POZ or BTB domain is also known as BR-C/Ttk or ZiN

[1] Zollman S, Godt D, Prive GG, Couderc JL, Laski FA; Proc Natl Acad Sci U S A 1994;91:10717-10721.

[2]Bardwell VJ, Treisman R; Genes Dev 1994;8:1664-1677.

30 [3] Ahmad KF, Engel CK, Prive GG; Proc Natl Acad Sci U S A 1998;95:12123-12128.

[4] Deweindt C, Albagli O, Bernardin F, Dhordain P, Quief S, Lantoine D, Kerckaert JP, Leprince D; Cell Growth Differ 1995;6:1495-1503.

[5] Huynh KD, Bardwell VJ; Oncogene 1998;17:2473-2484.

[6] Wong CW, Privalsky ML; J Biol Chem 1998;273:27695-27702.

62. (Bac GSPproteins) Bacterial type II secretion system protein D signature

5 A number of bacterial proteins, some of which are involved in a general secretion pathway (GSP) for the export of proteins (also called the type II pathway) [1 to 5], have been found to be evolutionary related. These proteins are listed below: - The 'D' protein from the GSP operon of: *Aeromonas* (gene *exdD*); *Erwinia* (gene *outD*); *Escherichia coli* (gene *yheF*), *Klebsiella pneumoniae* (gene *pulD*); *Pseudomonas aeruginosa* (gene *xcpQ*); *Vibrio cholerae* 10 (gene *epsD*) and *Xanthomonas campestris* (gene *xpsD*). - *comE* from *Haemophilus influenzae*, involved in competence (DNA uptake). - *pilQ* from *Pseudomonas aeruginosa*, which is essential for the formation of the pili. - *hofQ* (*hopQ*) from *Escherichia coli*. - *hrpH* from *Pseudomonas syringae*, which is involved in the secretion of a proteinaceous elicitor of the hypersensitivity response in plants. - *hrpA1* from *Xanthomonas campestris* pv. 15 *vesicatoria*, which is also involved in the hypersensitivity response. - *mxlD* from *Shigella flexneri* which is involved in the secretion of the *Ipa* invasins which are necessary for penetration of intestinal epithelial cells. - *omc* from *Neisseria gonorrhoeae*. - *yssC* from *Yersinia enterocolitica* virulence plasmid pYV, which seems to be required for the export of the Yop virulence proteins. - The gpIV protein from filamentous phages such as f1, ike, or 20 m13. GpIV is said to be involved in phage assembly and morphogenesis. These proteins all seem to start with a signal sequence and are thought to be integral proteins in the outer membrane. As a signature pattern a conserved region in the C-terminal section of these proteins has been selected

25 Consensus pattern: [GR]-[DEQKG]-[STVM]-[LIVMA](3)-[GA]-G-[LIVMFY]-x(11)-[LIVM]-P-[LIVMFYWGS]-[LIVMF]-[GSAE]-x-[LIVM]-P- [LIVMFYW](2)-x(2)-[LV]-F

[1] Salmond G.P.C., Reeves P.J. Trends Biochem. Sci. 18:7-12(1993).

[2] Reeves P.J., Whitcombe D., Wharam S., Gibson M., Allison G., Bunce N., Barallon R., 30 Douglas P., Mulholland V., Stevens S., Walker S., Salmond G.P.C. Mol. Microbiol. 8:443-456(1993).

[3] Martin P.R., Hobbs M., Free P.D., Jeske Y., Mattick J.S. Mol. Microbiol. 9:857-868(1993).

[4] Hobbs M., Mattick J.S. Mol. Microbiol. 10:233-243(1993).

[5] Genin S., Boucher C.A. Mol. Gen. Genet. 243:112-118(1994).

5 63. (Bac globin) Protozoan/cyanobacterial globins signature

Globins are heme-containing proteins involved in binding and/or transporting oxygen [1]. Almost all globins belong to a large family (see <PDOC00793>), the only exceptions are the following proteins which form a family of their own[2,3]: - Monomeric hemoglobins from the protozoan *Paramecium caudatum*, *Tetrahymena pyriformis* and *Tetrahymena*
10 *thermophila*. - Cyanoglobin from the cyanobacteria *Nostoc commune*. - Globins LI637 and LI410 from the chloroplast of the alga *Chlamydomonas eugametos*. - *Mycobacterium tuberculosis* hypothetical protein MtCY48.23. These proteins contain a conserved histidine which could be involved in heme-binding. As a signature pattern, a conserved region that ends with this residue was used

15 Consensus pattern: F-[LF]-x(5)-G-[PA]-x(4)-G-[KRA]-x-[LIVM]-x(3)-H-

[1] Concise Encyclopedia Biochemistry, Second Edition, Walter de Gruyter, Berlin New-York (1988).

20 [2] Takagi T. Curr. Opin. Struct. Biol. 3:413-418(1993).

[3] Couture M., Chamberland H., St-Pierre B., Lafontaine J., Guertin M.; Mol. Gen. Genet. 243:185-197(1994).

25 64. Band 7 protein family signature

Mammalian band 7 protein [1] (also known as 7.2B or stomatin) is an integral membrane phosphoprotein of red blood cells thought to regulate cation conductance by interacting with other proteins of the junctional complex of the membrane skeleton. Structurally, band 7 is evolutionary related to the following proteins: - *Caenorhabditis elegans* protein mec-2 [2].
30 Mec-2 positively regulates the activity of the putative mechanosensory transduction channel. It may links the mechanosensory channel and the microtubule cytoskeleton of the touch receptor neurons. - *Caenorhabditis elegans* proteins sto-1 to sto-4. - *Caenorhabditis elegans* protein unc-1. - *Escherichia coli* hypothetical protein ybbK. - *Mycobacterium tuberculosis*

hypothetical protein MtCY277.09. - Synechocystis strain PCC 6803 hypothetical protein slr1128. - Methanococcus jannaschii hypothetical protein MJ0827. Structurally all these proteins consist of a short N-terminal domain which is followed by a transmembrane region and a variable size (from 170 to 350 residues) C-terminal domain. As a signature pattern, a conserved region located about 110 residues after the transmembrane domain was selected

Consensus pattern: R-x(2)-[LIV]-[SAN]-x(6)-[LIV]-D-x(2)-T-x(2)-W-G-[LIV]-[KRH]-[LIV]-x-[KR]-[LIV]-E-[LIV]-[KR]-

[1] Gallagher P.G., Forget B.G. *J. Biol. Chem.* 270:26358-26363(1995).
[2] Huang M., Gu G., Ferguson E.L., Chalfie M. *Nature* 378:292-295(1995).

65. Barwin domain signatures

Barwin [1] is a barley seed protein of 125 residues that binds weakly a chitin analog. It contains six cysteines involved in disulfide bonds, as shown in the following schematic representation.

```
+-----+ | ***** | ****
xxxxxxxxxxxxxxxxCxxxxxxxxCxxxxCxCxxxxxxxxCxxxxxxxxxxxxxxxxCx || | +-----
-----+ +-----+ 'C': conserved cysteine involved in a disulfide bond. '*':
```

position of the patterns. Barwin is closely related to the following proteins: - Hevein, a wound-induced protein found in the latex of rubber trees. - HEL, an Arabidopsis thaliana hevein-like protein [2]. - Win1 and win2, two wound-induced proteins from potato. - Pathogenesis-related protein 4 from tobacco. Hevein and the win1/2 proteins consist of an N-terminal chitin-binding domain followed by a barwin-like C-terminal domain. Barwin and its related proteins could be involved in a defense mechanism in plants. As signature patterns, two highly conserved regions that contain some of the cysteines were selected

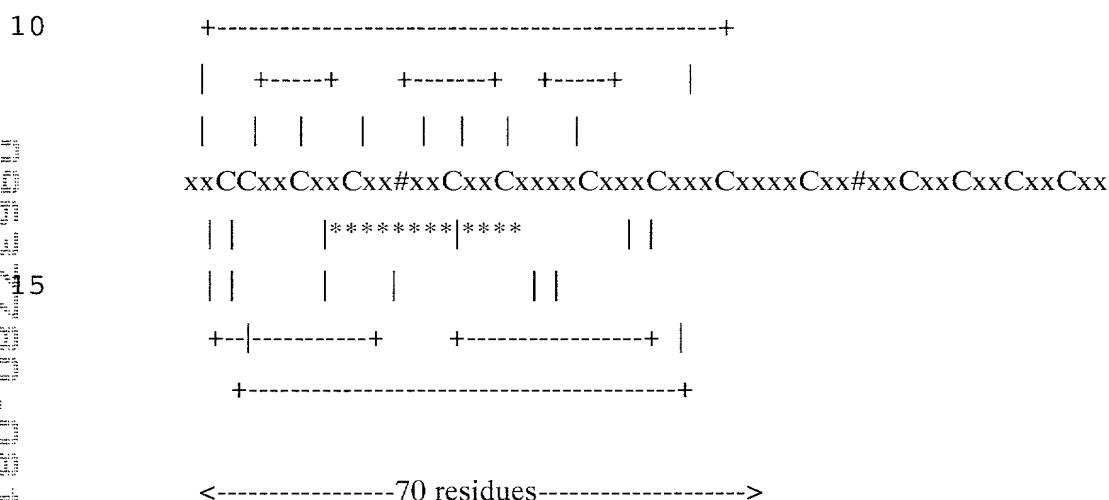
Consensus pattern: C-G-[KR]-C-L-x-V-x-N [The two C's are involved in disulfide bonds]-

Consensus pattern: V-[DN]-Y-[EQ]-F-V-[DN]-C [C is involved in a disulfide bond]-

[1] Svensson B., Svendsen I., Hoejrup P., Roepstorff P., Ludvigsen S., Poulsen F.M. *Biochemistry* 31:8767-8770(1992).

[2] Potter S., Uknes S., Lawton K., Winter A.M., Chandler D., Dimaio J., Novitzky R., Ward E., Ryals J. Mol. Plant Microbe Interact. 6:680-685(1993).

- 5 66. (Bowman-Birk leg) Bowman-Birk serine protease inhibitors family signature
PROSITE cross-reference(s). The Bowman-Birk inhibitor family [1] is one of the numerous families of serine proteinase inhibitors. As it can be seen in the schematic representation, they have a duplicated structure and generally possess two distinct inhibitory sites:



'C': conserved cysteine involved in a disulfide bond.

'#': active site residue.

15 '*': position of the pattern.

- 25 These inhibitors are found in the seeds of all leguminous plants as well as in cereal grains. In cereals they exist in two forms, one of which is a duplication of the basic structure shown above [2]. The pattern that was developed to pick up sequences belonging to this family of inhibitors is in the central part of the domain and includes four cysteines.

30 Consensus pattern C-x(5,6)-[DENQKRHSTA]-C-[PASTDH]-[PASTDK]-[ASTDV]-C-[NDKS]-[DEKRHSTA]-C [The four C's are involved in disulfide bonds] Note this pattern can be found twice in some duplicated cereal inhibitors.

[1] Laskowski M., Kato I. Annu. Rev. Biochem. 49:593-626(1980).

[2] Tashiro M., Hashino K., Shiozaki M., Ibuki F., Maki Z. J. Biochem. 102:297-306(1987).

5

67. Pathogenesis-related protein Bet v I family signature

A number of plant proteins, which all seem to be involved in pathogen defense response, are structurally related [1,2,3]. These proteins are:

- Bet v I, the major pollen allergen from white birch. Bet v I is the main cause of type I allergic reactions in Europe, North America and USSR.
- Aln g I, the major pollen allergen from alder.
- Api G I, the major allergen from celery.
- Car b I, the major pollen allergen from hornbeam.
- Cor a I, the major pollen allergen from hazel.
- Mal d I, the major pollen allergen from apple.
- Asparagus wound-induced protein AoPR1.
- Kidney bean pathogenesis-related proteins 1 and 2.
- Parsley pathogenesis-related proteins PR1-1 and PR1-3.
- Pea disease resistance response proteins pI49, pI176 and DRRG49-C.
- Pea abscisic acid-responsive proteins ABR17 and ABR18.
- Potato pathogenesis-related proteins STH-2 and STH-21.
- Soybean stress-induced protein SAM22.

These proteins are thought to be intracellularly located. They contain from 155 to 160 amino acid residues. As a signature pattern, a conserved region located in the third quarter of these proteins has been selected

Consensus pattern: G-x(2)-[LIVMF]-x(4)-E-x(2)-[CSTAEN]-x(8,9)-[GND]-G-[GS]- [CS]-x(2)-K-x(4)-[FY]-

[1] Breiteneder H., Pettenburger K., Bito A., Valenta R., Kraft D., Rumpold H., Scheiner O., Breitenbach M. EMBO J. 8:1935-1938(1989).

[2] Crowell D., John M.E., Russell D., Amasino R.M. Plant Mol. Biol. 18:459-466(1992).

[3] Warner S.A.J., Scott R., Draper J. Plant Mol. Biol. 19:555-561(1992).

68. bZIP transcription factors basic domain signature

The bZIP superfamily [1,2,] of eukaryotic DNA-binding transcription factors groups together proteins that contain a basic region mediating sequence-specific DNA-binding followed by a leucine zipper required for dimerization. This family is quite large, therefore only a partial list of some representative members appears here. - Transcription factor AP-1, which binds selectively to enhancer elements in the cis control regions of SV40 and metallothionein IIA. AP-1, also known as c-jun, is the cellular homolog of the avian sarcoma virus 17 (ASV17) oncogene v-jun. - Jun-B and jun-D, probable transcription factors which are highly similar to jun/AP-1. - The fos protein, a proto-oncogene that forms a non-covalent dimer with c-jun. - The fos-related proteins fra-1, and fos B. - Mammalian cAMP response element (CRE) binding proteins CREB, CREM, ATF-1, ATF-3, ATF-4, ATF-5, ATF-6 and LRF-1. - Maize Opaque 2, a trans-acting transcriptional activator involved in the regulation of the production of zein proteins during endosperm. - Arabidopsis G-box binding factors GBF1 to GBF4, Parsley CPRF-1 to CPRF-3, Tobacco TAF-1 and wheat EMBP-1. All these proteins bind the G-box promoter elements of many plant genes. - Drosophila protein Giant, which represses the expression of both the kruppel and knirps segmentation gap genes. - Drosophila Box B binding factor 2 (BBF-2), a transcriptional activator that binds to fat body-specific enhancers of alcohol dehydrogenase and yolk protein genes. - Drosophila segmentation protein cap'n'collar (gene cnc), which is involved in head morphogenesis. - Caenorhabditis elegans skn-1, a developmental protein involved in the fate of ventral blastomeres in the early embryo. - Yeast GCN4 transcription factor, a component of the general control system that regulates the expression of amino acid-synthesizing enzymes in response to amino acid starvation, and the related Neurospora crassa cpc-1 protein. - Neurospora crassa cys-3 which turns on the expression of structural genes which encode sulfur-catabolic enzymes. - Yeast MET28, a transcriptional activator of sulfur amino acids metabolism. - Yeast PDR4 (or YAP1), a transcriptional activator of the genes for some oxygen detoxification enzymes. - Epstein-Barr virus trans-activator protein BZLF1.-

Consensus pattern: [KR]-x(1,3)-[RKSAQ]-N-x(2)-[SAQ](2)-x-[RKTAENQ]-x-R-x-[RK]-

[1] Hurst H.C. Protein Prof. 2:105-168(1995).[2] Ellenberger T. Curr. Opin. Struct. Biol. 4:12-21(1994).

69. Biotin-requiring enzymes attachment site

Biotin, which plays a catalytic role in some carboxyl transfer reactions, is covalently attached, via an amide bond, to a lysine residue in enzymes requiring this coenzyme [1,2,3,4]. Such enzymes are:

- Pyruvate carboxylase (EC 6.4.1.1).
- Acetyl-CoA carboxylase (EC 6.4.1.2).
- Propionyl-CoA carboxylase (EC 6.4.1.3).
- Methylcrotonoyl-CoA carboxylase (EC 6.4.1.4).
- Geranoyl-CoA carboxylase (EC 6.4.1.5).
- Urea carboxylase (EC 6.3.4.6).
- Oxaloacetate decarboxylase (EC 4.1.1.3).
- Methylmalonyl-CoA decarboxylase (EC 4.1.1.41).
- Glutaconyl-CoA decarboxylase (EC 4.1.1.70).
- Methylmalonyl-CoA carboxyl-transferase (EC 2.1.3.1) (transcarboxylase).

Sequence data reveal that the region around the biocytin (biotin-lysine) residue is well conserved and can be used as a signature pattern.

Consensus pattern[GN]-[DEQTR]-x-[LIVMFY]-x(2)-[LIVM]-x-[AIV]-M-K-[LMAT]-x(3)-[LIVM]-x-[SAV] [K is the biotin attachment site] Note the domain around the biotin-binding lysine residue is evolutionary related to that around the lipoyl-binding lysine residue of 2-oxo acid dehydrogenase acyltransferases

- [1] Knowles J.R. Annu. Rev. Biochem. 58:195-221(1989).
- [2] Samols D., Thronton C.G., Murtif V.L., Kumar G.K., Haase F.C., Wood H.G. J. Biol. Chem. 263:6461-6464(1988).
- [3] Goss N.H., Wood H.G. Meth. Enzymol. 107:261-278(1984).
- [4] Shenoy B.C., Xie Y., Park V.L., Kumar G.K., Beegen H., Wood H.G., Samols D. J. Biol. Chem. 267:18407-18412(1992).

2-oxo acid dehydrogenases acyltransferase component lipoyl binding site

The 2-oxo acid dehydrogenase multienzyme complexes [1,2] from bacterial and

eukaryotic sources catalyze the oxidative decarboxylation of 2-oxo acids to the corresponding acyl-CoA. The three members of this family of multienzyme complexes are:

- Pyruvate dehydrogenase complex (PDC).
- 5 - 2-oxoglutarate dehydrogenase complex (OGDC).
- Branched-chain 2-oxo acid dehydrogenase complex (BCOADC).

These three complexes share a common architecture: they are composed of multiple copies of three component enzymes - E1, E2 and E3. E1 is a thiamine pyrophosphate-dependent 2-oxo acid dehydrogenase, E2 a dihydrolipamide acyltransferase, and E3 an FAD-containing dihydrolipamide dehydrogenase.

E2 acyltransferases have an essential cofactor, lipoic acid, which is covalently bound via an amide linkage to a lysine group. The E2 components of OGDC and BCOADC bind a single lipoyl group, while those of PDC bind either one (in yeast and in *Bacillus*), two (in mammals), or three (in *Azotobacter* and in *Escherichia coli*) lipoyl groups [3].

In addition to the E2 components of the three enzymatic complexes described above, a lipoic acid cofactor is also found in the following proteins:

- H-protein of the glycine cleavage system (GCS) [4]. GCS is a multienzyme complex of four protein components, which catalyzes the degradation of glycine. H protein shuttles the methylamine group of glycine from the P protein to the T protein. H-protein from either prokaryotes or eukaryotes binds a single lipoic group.
- Mammalian and yeast pyruvate dehydrogenase complexes differ from that of other sources, in that they contain, in small amounts, a protein of unknown function - designated protein X or component X. Its sequence is closely related to that of E2 subunits and seems to bind a lipoic group [5].
- Fast migrating protein (FMP) (gene *acoC*) from *Alcaligenes eutrophus* [6]. This protein is most probably a dihydrolipamide acyltransferase involved in acetoin metabolism.

A signature pattern was developed which allows the detection of the lipoyl-binding site.

Consensus pattern[GN]-x(2)-[LIVF]-x(5)-[LIVFC]-x(2)-[LIVFA]-x(3)-K-[STAIV]-
[STAVQDN]-x(2)-[LIVMFS]-x(5)-[GCN]-x-[LIVMFY] [K is the lipoyl-binding site] Note
the domain around the lipoyl-binding lysine residue is evolutionary related to that around the
biotin-binding lysine residue of biotin requiring enzymes

5

- [1] Yeaman S.J. Biochem. J. 257:625-632(1989).
- [2] Yeaman S.J. Trends Biochem. Sci. 11:293-296(1986).
- [3] Russel G.C., Guest J.R. Biochim. Biophys. Acta 1076:225-232(1991).
- [4] Fujiwara K., Okamura-Ikeda K., Motokawa Y. J. Biol. Chem. 261:8836-8841(1986).
- 10 [5] Behal R.H., Browning K.S., Hall T.B., Reed L.J. Proc. Natl. Acad. Sci. U.S.A. 86:8732-8736(1989).
- [6] Priefert H., Hein S., Krueger N., Zeh K., Schmidt B., Steinbuechel A. J. Bacteriol. 173:4056-4071(1991).

15
20

70. C2 (C2 domain) Number of members: 295

Some isozymes of protein kinase C (PKC) [1,2] contain a domain, known as C2, of about 116 amino-acid residues which is located between the two copies of the C1 domain (that bind phorbol esters and diacylglycerol) (see <PDOC00379>) and the protein kinase catalytic domain (see <PDOC00100>). Regions with significant homology [3,E1] to the C2-domain have been found in the following proteins:

- PKC isoforms alpha, beta and gamma and Drosophila isoforms PKC1 and PKC2.
- PKC isoforms delta, epsilon and eta, Caenorhabditis elegans kin-13 and yeast PKC1 have a C2-like domain at the N-terminal extremity [4].
- 25 - Yeast cAMP dependent protein kinase SCH9 contains a C2-like domain.
- Mammalian phosphatidylinositol-specific phospholipase C (PI-PLC) (see <PDOC50007>) isoforms beta, gamma and delta as well as several non-mammalian PI-PLCs have a C2-like domain C-terminal of the catalytic domain.
- Mammalian and plants phosphatidylinositol-3-kinase have a C2-like domain in the central
- 30 region of the 110 Kd catalytic subunit.
- Yeast phosphatidylserine-decarboxylase 2 (gene PSD2) contains a C2 domain in its central region.

- Cytosolic phospholipase D from plants and cytosolic phospholipase A2 have a C2-like domain at their N-terminus.

- Synaptotagmins (p65). This is a family of related synaptic vesicle proteins that bind acidic phospholipids and that may have a regulatory role in the membrane interactions during trafficking of synaptic vesicles at the active zone of the synapse. All isoforms of synaptotagmins have two copies of the C2 domain in their C-terminal region.

- Rabphilin-3A, a synaptic protein contains two C2 domains.

- *Caenorhabditis elegans* protein unc-13 whose function is not known. Unc-13 has a C2 domain in its central part and a C2-like domain at the C-terminus.

- rasGAP and the breakpoint cluster protein bcr have a C2-domain C-terminal of a PH-domain.

- Yeast protein BUD2 (or CLA2) has a C2-domain in the central region.

- Yeast protein RSP5 and human protein NEDD-4, both proteins also contain WW domains (see <PDOC50020>).

- Perforin (see <PDOC00251>) has a C2 domain at the C-terminus. It is the only extracellular protein known to contain a C2 domain.

- Yeast hypothetical protein YML072C has a C2 domain.

- Yeast hypothetical protein YNL087W has three C2 domains.

- *Caenorhabditis elegans* hypothetical protein F37A4.7 has two C2 domains.

The C2 domain is thought to be involved in calcium-dependent phospholipid binding [5]. Since domains related to the C2 domain are also found in proteins that do not bind calcium, other putative functions for the C2 domain like e.g. binding to inositol-1,3,4,5-tetrakisphosphate have been suggested [6]. Recently, the 3D structure of the first C2 domain of synaptotagmin has been reported [7], the domain forms an eight-stranded beta sandwich. The signature pattern that has been developed for the C2 domain is located in a conserved part of that domain, the connecting loop between beta strands 2 and 3. A profile has been developed for the C2 domain that covers the total domain.

-Consensus pattern: [ACG]-x(2)-L-x(2,3)-D-x(1,2)-[NGSTLIF]-[GTMR]-x-[STAP]-D-[PA]-[FY]

-Note: this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so.

[1]Medline: 96367095 Extending the C2 domain family: C2s in PKCs delta, epsilon, eta and theta, phospholipases, GAPs and perforin. Ponting CP, Parker PJ; Protein Sci 1996;5:162-166.

- 5 [1] Azzi A., Boscoboinik D., Hensey C. Eur. J. Biochem. 208:547-557(1992).
- [2] Stabel S. Semin. Cancer Biol. 5:277-284(1994).
- [3] Brose N., Hofmann K.O., Hata Y., Suedhof T.C. J. Biol. Chem. 270:25273-25280(1995).
- [4] Sossin W.S., Schwartz J.H. Trends Biochem. Sci. 18:207-208(1993).
- [5] Davletov B.A., Suedhof T.C. J. Biol. Chem. 268:26386-26390(1993).
- 10 [6] Fukuda M., Aruga J., Niinobe M., Aimoto S., Mikoshiba K. J. Biol. Chem. 269:29206-29211(1994).
- [6] Sutton R.B., Davletov B.A., Berghuis A.M., Suedhof T.C., Sprang S.R. Cell 80:929-938(1995).

71. CAP (CAP protein) Number of members: 11

In budding and fission yeasts the CAP protein is a bifunctional protein whose N-terminal domain binds to adenylyl cyclase, thereby enabling that enzyme to be activated by upstream regulatory signals, such as Ras. The function of the C-terminal domain is less clear, but it is required for normal cellular morphology and growth control [1]. CAP is conserved in higher eukaryotic organisms where its function is not yet clear [2].

Structurally, CAP is a protein of 474 to 551 residues which consist of two domains separated by a proline-rich hinge. Two signature patterns, one corresponding to a conserved region in the N-terminal extremity and the other to a C-terminal region have been developed.

-Consensus pattern: [LIVM](2)-x-R-L-[DE]-x(4)-R-L-E

-Consensus pattern: D-[LIVMFY]-x-E-x-[PA]-x-P-E-Q-[LIVMFY]-K

[1] Kawamukai M., Gerst J., Field J., Riggs M., Rodgers L., Wigler M., Young D. Mol. Biol. Cell 3:167-180(1992).

[2] Yu G., Swiston J., Young D. J. Cell Sci. 107:1671-1678(1994).

72. CAP_GLY (CAP-Gly domain)

CAP stands for cytoskeleton-associated proteins. Swiss:P39937 may be a member but has not been included. It has a weak match to the family between residues 22-67. Number of members: 24

5

[1]Medline: 93242656. Sequence homologies between four cytoskeleton-associated proteins. Riehemann K, Sorg C; Trends Biochem Sci 1993;18:82-83.

10

It has been shown [1] that some cytoskeleton-associated proteins (CAP) share the presence of a conserved, glycine-rich domain of about 42 residues, called here CAP-Gly. Proteins known to contain this domain are listed below.

15

- Restin (also known as cytoplasmic linker protein-170 or CLIP-170), a 160 Kd protein associated with intermediate filaments and that links endocytic vesicles to microtubules. Restin contains two copies of the CAP-Gly domain.

20

- Vertebrate dynactin (150 Kd dynein-associated polypeptide; DAP) and Drosophila glued, a major component of activator I, a 20S polypeptide complex that stimulates dynein-mediated vesicle transport.

- Yeast protein BIK1 which seems to be required for the formation or stabilization of microtubules during mitosis and for spindle pole body fusion during conjugation.

- Yeast protein NIP100 (NIP80).

- Human protein CKAP1/TFCB, Schizosaccharomyces pombe protein alp11 and Caenorhabditis elegans hypothetical protein F53F4.3. These proteins contain a N-terminal ubiquitin domain (see <PDOC00271>) and a C-terminal CAP-Gly domain.

- Caenorhabditis elegans hypothetical protein M01A8.2.

25

- Yeast hypothetical protein YNL148c.

Structurally, these proteins are made of three distinct parts: an N-terminal section that is most probably globular and contains the CAP-Gly domain, a large central region predicted to be in an alpha-helical coiled-coil conformation and, finally, a short C-terminal globular domain. The signature for the CAP-Gly domain corresponds to the first 32 residues of the domain and includes five of the six conserved glycines.

30

-Consensus pattern: G-x(8,10)-[FYW]-x-G-[LIVM]-x-[LIVMFY]-x(4)-G-K-[NH]-x-G-[STAR]-x(2)-G-x(2)-[LY]-F

[1] Riehemann K., Sorg C. Trends Biochem. Sci. 18:82-83(1993).

5 73. (CBD 1)

Cellulose-binding domain, fungal type

The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1].

10

Structurally, cellulases and xylanases generally consist of a catalytic domain joined to a cellulose-binding domain (CBD) by a short linker sequence rich in proline and/or hydroxy-amino acids.

15

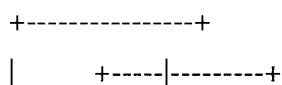
The CBD of a number of fungal cellulases has been shown to consist of 36 amino acid residues. Enzymes known to contain such a domain are:

20

- Endoglucanase I (gene egl1) from *Trichoderma reesei*.
- Endoglucanase II (gene egl2) from *Trichoderma reesei*.
- Endoglucanase V (gene egl5) from *Trichoderma reesei*.
- Exocellobiohydrolase I (gene CBHI) from *Humicola grisea*, *Neurospora crassa*, *Phanerochaete chrysosporium*, *Trichoderma reesei*, and *Trichoderma viride*.
- Exocellobiohydrolase II (gene CBHII) from *Trichoderma reesei*.
- Exocellobiohydrolase 3 (gene cel3) from *Agaricus bisporus*
- 25 - Endoglucanases B, C2, F and K from *Fusarium oxysporum*.

30

The CBD domain is found either at the N-terminal (Cbh-II or egl2) or at the C-terminal extremity (Cbh-I, egl1 or egl5) of these enzymes. As it is shown in the following schematic representation, there are four conserved cysteines in this type of CBD domain, all involved in disulfide bonds.




```

      |   |   |   |
xxxxxxx Cxxxxxxxxxx Cxxxxx Cxxxxxxxxxx Cx
*****

```

- 5 'C': conserved cysteine involved in a disulfide bond.
 '*': position of the pattern.

10 Such a domain has also been found in a putative polysaccharide binding protein from the red alga, *Porphyra purpurea* [2]. Structurally, this protein consists of four tandem repeats of the CBD domain.

Consensus pattern C-G-G-x(4,7)-G-x(3)-C-x(5)-C-x(3,5)-[NHG]-x-[FYWM]-x(2)-Q-C [The four C's are involved in disulfide bonds] Sequences known to belong to this class detected by the pattern ALL.

[1] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).

[2] Liu Q., der Meer J.P., Reith M.E.

20 74. CBS domain. 3D Structure found as a subdomain in TIM barrel of inosine-. CBS domain web page. CBS domains are small intracellular modules mostly found in 2 or four copies within a protein. CBS domains are found in cystathionine-beta-synthase (CBS) where mutations lead to homocystinuria. Two CBS domains are found in inosine-monophosphate dehydrogenase from all species, however the CBS domains are not needed for activity. Two CBS domains are found in intracellular loops of several chloride channels. Mutations in this domain of Swiss:P35520 lead to homocystinuria.

Number of members: 414

- 30 [1]Medline: 97172695 The structure of a domain common to archaeobacteria and the homocystinuria disease protein. Bateman A; Trends Biochem Sci 1997;22:12-13.
 [2]Medline: 96279836 Structure and mechanism of inosine monophosphate dehydrogenase in complex with the immunosuppressant mycophenolic-acid. Sintchak MD, Fleming MA,

Futer O, Raybuck SA, Chambers SP, Caron PR, Murcko MA, Wilson KP; Cell 1996;85:921-930.

Discovery of CBS domain.

[3]Medline: 97259972 CBS domains in ClC chloride channels implicated in myotonia and nephrolithiasis (kidney stones). Ponting CP; J Mol Med 1997;75:160-163.

75. CDP-OH_P_transf (CDP-alcohol phosphatidyltransferase)

All of these members have the ability to catalyze the displacement of CMP from a CDP-alcohol by a second alcohol with formation of a phosphodiester bond and concomitant breaking of a phosphoride anhydride bond. Number of members: 32

A number of phosphatidyltransferases, which are all involved in phospholipid biosynthesis and that share the property of catalyzing the displacement of CMP from a CDP-alcohol by a second alcohol with formation of a phosphodiester bond and concomitant breaking of a phosphoride anhydride bond share a conserved sequence region [1,2]. These enzymes are:

- Ethanolaminephosphotransferase (EC 2.7.8.1) from yeast (gene EPT1).
- Diacylglycerol cholinephosphotransferase (EC 2.7.8.2) from yeast (gene CPT1).
- Phosphatidylglycerophosphate synthase (EC 2.7.8.5) (CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase) from bacteria (gene pgsA).
- Phosphatidylserine synthase (EC 2.7.8.8) (CDP-diacylglycerol--serine O-phosphatidyltransferase) from yeast (gene CHO1) and from *Bacillus subtilis* (gene pssA).
- Phosphatidylinositol synthase (EC 2.7.8.11) (CDP-diacylglycerol--inositol 3-phosphatidyltransferase) from yeast (gene PIS).

These enzymes are proteins of from 200 to 400 amino acid residues. The conserved region contains three aspartic acid residues and is located in the N-terminal section of the sequences.

-Consensus pattern: D-G-x(2)-A-R-x(8)-G-x(3)-D-x(3)-D

[1]Medline: 97075020 Two-dimensional 1H-NMR of transmembrane peptides from *Escherichia coli* phosphatidylglycerophosphate synthase in micelles. Morein S, Trouard TP, Hauksson JB, Rilfors L, Arvidson G, Lindblom G; Eur J Biochem 1996;241:489-497.

[1] Nikawa J.-I., Kodaki T., Yamashita S.

J. Biol. Chem. 262:4876-4881(1987).

[2] Hjelmstad R.H., Bell R.M.

J. Biol. Chem. 266:5094-5134(1991).

5

76. CHOD (Cholesterol oxidase) Members of the GMC oxidoreductase family. Number of members: 3

10

[1]Medline: 94032271. Crystal structure of cholesterol oxidase complexed with a steroid substrate: implications for flavin adenine dinucleotide dependent alcohol oxidases. Li J, Vrielink A, Brick P, Blow DM; Biochemistry 1993;32:11507-11515.

11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

The following FAD flavoproteins oxidoreductases have been found [1,2] to be evolutionary related. These enzymes, which are called 'GMC oxidoreductases', are listed below.

- Glucose oxidase (EC 1.1.3.4) (GOX) from *Aspergillus niger*. Reaction catalyzed: glucose + oxygen -> delta-luconolactone + hydrogen peroxide.
 - Methanol oxidase (EC 1.1.3.13) (MOX) from fungi. Reaction catalyzed: methanol + oxygen -> acetaldehyde + hydrogen peroxide.
 - Choline dehydrogenase (EC 1.1.99.1) (CHD) from bacteria. Reaction catalyzed: choline + unknown acceptor -> betaine acetaldehyde + reduced acceptor.
 - Glucose dehydrogenase (GLD) (EC 1.1.99.10) from *Drosophila*. Reaction catalyzed: glucose + unknown acceptor -> delta-gluconolactone + reduced acceptor.
 - Cholesterol oxidase (CHOD) (EC 1.1.3.6) from *Brevibacterium sterolicum* and *Streptomyces* strain SA-COO. Reaction catalyzed: cholesterol + oxygen -> cholest-4-en-3-one + hydrogen peroxide.
 - AlkJ [3], an alcohol dehydrogenase from *Pseudomonas oleovorans*, which converts aliphatic medium-chain-length alcohols into aldehydes. This family also includes a lyase:
 - (R)-mandelonitrile lyase (EC 4.1.2.10) (hydroxynitrile lyase) from plants [4], an enzyme involved in cyanogenesis, the release of hydrogen cyanide from injured tissues.
- These enzymes are proteins of size ranging from 556 (CHD) to 664 (MOX) amino acid residues which share a number of regions of sequence similarities. One of these regions, located in the N-terminal section, corresponds to the FAD ADP- binding domain. The function of the other conserved domains is not yet known; two of these domains have been

selected as signature patterns. The first one is located in the N-terminal section of these enzymes, about 50 residues after the ADP-binding domain, while the second one is located in the central section.

5 -Consensus pattern: [GA]-[RKN]-x-[LIV]-G(2)-[GST](2)-x-[LIVM]-N-x(3)-[FYWA]-x(2)-[PAG]-x(5)-[DNESH]

-Consensus pattern: [GS]-[PSTA]-x(2)-[ST]-P-x-[LIVM](2)-x(2)-S-G-[LIVM]-G

[1] Cavener D.R. J. Mol. Biol. 223:811-814(1992).

10 [2] Henikoff S., Henikoff J.G. Genomics 19:97-107(1994).

[3] van Beilen J.B., Eggink G., Enequist H., Bos R., Witholt B. Mol. Microbiol. 6:3121-3136(1992).

[4] Cheng I.P., Poulton J.E. Plant Cell Physiol. 34:1139-1143(1993).

15 77. CKS (Cyclin-dependent kinase regulatory subunit) Number of members: 11. Cyclin-dependent kinases (CDK) are protein kinases which associate with cyclins to regulate eukaryotic cell cycle progression. The most well known CDK is p34-cdc2 (CDC28 in yeast) which is required for entry into S-phase and mitosis. CDK's bind to a regulatory subunit which is essential for their biological function. This regulatory subunit is a small protein of 20 79 to 150 residues. In yeast (gene CKS1) and in fission yeast (gene suc1) a single isoform is known, while mammals have two highly related isoforms. It has been shown [1] that these CDK regulatory subunits assemble as a hexamer which then acts as a hub for the oligomerization of six CDK catalytic subunits. The sequence of CDK regulatory subunits are 25 highly conserved therefore, the two most conserved regions have been used as signature patterns.

-Consensus pattern: Y-S-x-[KR]-Y-x-[DE](2)-x-[FY]-E-Y-R-H-V-x-[LV]-[PT]-[KRP]

-Consensus pattern: H-x-P-E-x-H-[IV]-L-L-F-[KR]

[1] Parge H.E., Arvai A.S., Murtari D.J., Reed S.I., Tainer J.A. Science 262:387-395(1993).

78. CK_II_beta (Casein kinase II regulatory subunit)

Number of members: 16. Casein kinase II (CK-2) [1] is an ubiquitous eukaryotic serine/threonine protein kinase which is found both in the cytoplasm and the nucleus and whose substrates are numerous. It generally phosphorylates Ser or Thr at the N-terminal of stretch of acidic residues (see <PDOC00006>). CK-2 exists as an heterotetramer composed of two catalytic subunits (alpha) and two regulatory subunits (beta). In most species there are two closely related isoforms of the catalytic subunit: alpha and alpha'. Some species, such as fungi and plants, express two forms of regulatory subunits: beta and beta'. The exact function of the regulatory subunit is not yet known. It is a highly conserved protein of about 25 Kd that contains, in its central section, a cysteine-rich motif that could be involved in binding a metal such as zinc [2]. This region has been used as a signature pattern.

-Consensus pattern: C-P-x-[LIVMY]-x-C-x(5)-[LI]-P-[LIVMC]-G-x(9)-V-[KR]-x(2)-C-P-x-C

[1] Allende J.E., Allende C.C. FASEB J. 9:313-323(1995).

[2] Reed J.C., Bidwai A.P., Glover C.V.C. J. Biol. Chem. 269:18192-18200(1994).

79. CLP_protease (Clp protease)

These proteins belong to family S14 in the classification of peptidases.

-!- The Clp protease has an active site catalytic triad. In E. coli Clp protease, ser-111, his-136 and asp-185 form the catalytic triad.

-!- Swiss:P48254 has lost all of these active site residues and is therefore inactive.

-!- Swiss:P42379 contains two large insertions, Swiss:P42380 contains one large insertion.

Number of members: 38

The endopeptidase Clp (EC 3.4.21.92) from Escherichia coli cleaves peptides in various proteins in a process that requires ATP hydrolysis [1,2]. Clp is a dimeric protein which consists of a proteolytic subunit (gene clpP) and either of two related ATP-binding regulatory subunits (genes clpA and clpX). ClpP is a serine protease which has a chymotrypsin-like activity. Its catalytic activity seems to be provided by a charge relay system similar to that of the trypsin family of serine proteases, but which evolved by independent convergent

evolution. Proteases highly similar to ClpP have been found to be encoded in the genome of the chloroplast of plants and seem to be also present in other eukaryotes. The sequences around two of the residues involved in the catalytic triad (a serine and a histidine) are highly conserved and can be used as signature patterns specific to that category of proteases.

-Consensus pattern: T-x(2)-[LIVMF]-G-x-A-[SAC]-S-[MSA]-[PAG]-[STA] [S is the active site residue]

-Consensus pattern: R-x(3)-[EAP]-x(3)-[LIVMFYT]-M-[LIVM]-H-Q-P [H is the active site residue]

[1]Medline: 98050920. The structure of ClpP at 2.3 angstroms resolution suggests a model for ATP-dependent proteolysis. Wang J, Hartling JA, Flanagan JM; Cell 1997;91:447-456.

[1] Maurizi M.R., Clark W.P., Kim S.-H., Gottesman S. J. Biol. Chem. 265:12546-12552(1990).

[2] Gottesman S., Maurizi M.R. Microbiol. Rev. 56:592-621(1992).

[3] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

80. CNG_membrane (Transmembrane region cyclic Nucleotide Gated Channel)

[1]Medline: 94224763. Cyclic nucleotide-gated channels: an expanding new family of ion channels. Yau KW; Proc Natl Acad Sci USA 1994;91:3481-3483.

This family is found to the N-terminus of the cNMP_binding. Number of members: 56.

Proteins that bind cyclic nucleotides (cAMP or cGMP) share a structural domain of about 120 residues [1-3]. The best studied of these proteins is the prokaryotic catabolite gene activator (also known as the cAMP receptor protein) (gene crp) where such a domain is known to be composed of three alpha-helices and a distinctive eight-stranded, antiparallel beta-barrel structure. Such a domain is known to exist in the following proteins:

- Prokaryotic catabolite gene activator protein (CAP).

- cAMP- and cGMP-dependent protein kinases (cAPK and cGPK). Both types of kinases contains two tandem copies of the cyclic nucleotide-binding domain. The cAPK's are composed of two different subunits: a catalytic chain and a regulatory chain which contains both copies of the domain. The cGPK's are single chain enzymes that include the two copies

of the domain in their N-terminal section. The nucleotide specificity of cAPK and cGPK is due to an amino acid in the conserved region of beta-barrel 7: a threonine that is invariant in cGPK is an alanine in most cAPK.

- Vertebrate cyclic nucleotide-gated ion-channels. Two such cations channels have been

fully characterized. One is found in rod cells where it plays a role in visual signal transduction. It specifically binds to cGMP leading to an opening of the channel and thereby causing a depolarization of rod photoreceptors. In olfactory epithelium a similar, cAMP-binding, channel plays a role in odorant signal transduction. There are six invariant amino acids in this domain, three of which are glycine residues that are thought to be essential for maintenance of the structural integrity of the beta-barrel. Two signature patterns have been developed for this domain. The first pattern is located within beta-barrels and 3 and contains the first two conserved Gly. The second pattern is located within beta-barrels 6 and 7 and contains the third conserved Gly as well as the three other invariant residues.

-Consensus pattern: [LIVM]-[VIC]-x(2)-G-[DENQTA]-x-[GAC]-x(2)-[LIVMFY](4)-x(2)-G

-Consensus pattern: [LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV]

[1] Weber I.T., Shabb J.B., Corbin J.D. Biochemistry 28:6122-6127(1989).

[2] Kaupp U.B. Trends Neurosci. 14:150-157(1991).

[3] Shabb J.B., Corbin J.D. J. Biol. Chem. 267:5723-5726(1992).

81. COX10_ctaB_cyoE (Cytochrome c oxidase assembly factor)

[1]Medline: 95191390

Biosynthesis and functional role of haem O and haem A

Mogi T, Saiki K, Anraku Y; Mol Microbiol 1994;14:391-398.

Cytochrome c oxidase is a multi subunit enzyme. The complexity of this enzyme requires assistance in building the complex.

This is carried out by the Cytochrome c oxidase assembly factor.

Number of members: 31

Cytochrome c oxidase is an oligomeric enzymatic complex which seems to require the aid of a number of proteins that either act as chaperonins to help the subunits of the enzyme to fold correctly, or assist in the assembly of the metal centers [1]. One of these subunits is known as COX10 in yeast and as
 5 ctaB [2] in aerobic prokaryotes. It is evolutionary related to cyoE protein from the Escherichia coli cytochrome O terminal oxidase complex.

These proteins probably contain [3] seven transmembrane segments. The most conserved region is located in a loop between the second and third of these
 10 segments and has been selected as a signature pattern.

-Consensus pattern: [ED]-x-D-x(2)-M-x-R-T-x(2)-R-x(4)-G

[1] Nobrega M.P., Nobrega F.G., Tzagoloff A.

J. Biol. Chem. 265:14220-14226(1990).

[2] Cao J., Hosler J., Shapleigh J., Revzin A., Ferguson-Miller S.

J. Biol. Chem. 267:24273-24278(1992).

[3] Chepuri V., Gennis R.B.

J. Biol. Chem. 265:12978-12986(1990).

82. COX3 (Cytochrome c oxidase subunit III)

This family corresponds to chains c and p.

[1]Medline: 96216288

25 The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 Å. Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S; Science 1996;272:1136-1144.
 Number of members: 224

83. COX5B (Cytochrome c oxidase subunit Vb)

[1]

Medline: 96216288

The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 Å.

Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S;

Science 1996;272:1136-1144.

This family consists of chains F and S

Number of members: 10

Cytochrome c oxidase (EC 1.9.3.1) [1] is an oligomeric enzymatic complex which is a component of the respiratory chain complex and is involved in the transfer of electrons from cytochrome c to oxygen. In eukaryotes this enzyme complex is located in the mitochondrial inner membrane; in aerobic prokaryotes it is found in the plasma membrane. In addition to the three large subunits that form the catalytic center of the enzyme complex there are, in eukaryotes, a variable number of small polypeptidic subunits. One of these subunits which is known as Vb in mammals, V in slime mold and IV in yeast, binds a zinc atom. The sequence of subunit Vb is well conserved and includes three conserved cysteines that are thought to coordinate the zinc ion [2]. Two of these cysteines are clustered in the C-terminal section of the subunit; this region has been selected as a signature pattern.

-Consensus pattern: [LIVM](2)-[FYW]-x(10)-C-x(2)-C-G-x(2)-[FY]-K-L [The two C's probably bind zinc]

[1] Capaldi R.A., Malatesta F., Darley-USmar V.M.

Biochim. Biophys. Acta 726:135-148(1983).

[2] Rizzuto R., Sandona D., Brini M., Capaldi R.A., Bisson R.

Biochim. Biophys. Acta 1129:100-104(1991).

84. COesterase (Carboxylesterases)

Cholinesterase pages

The prints entry is specific to acetylcholinesterase

Number of members: 273

Higher eukaryotes have many distinct esterases. Among the different types are those which act on carboxylic esters (EC 3.1.1.-). Carboxyl-esterases have been classified into three categories (A, B and C) on the basis of differential patterns of inhibition by organophosphates. The sequence of a number of type-B carboxylesterases indicates [1,2,3] that the majority are evolutionary related. This family currently consists of the following proteins:

- Acetylcholinesterase (EC 3.1.1.7) (AChE) [E1] from vertebrates and from *Drosophila*.
- Mammalian cholinesterase II (butyryl cholinesterase) (EC 3.1.1.8).
Acetylcholinesterase and cholinesterase II are closely related enzymes that hydrolyze choline esters [4].
- Mammalian liver microsomal carboxylesterases (EC 3.1.1.1).
- *Drosophila* esterase 6, produced in the anterior ejaculatory duct of the male insect reproductive system where it plays an important role in its reproductive biology.
- *Drosophila* esterase P.
- *Culex pipiens* (mosquito) esterases B1 and B2.
- *Myzus persicae* (peach-potato aphid) esterases E4 and FE4.
- Mammalian bile-salt-activated lipase (BAL) [5], a multifunctional lipase which catalyzes fat and vitamin absorption. It is activated by bile salts in infant intestine where it helps to digest milk fats.
- Insect juvenile hormone esterase (JH esterase) (EC 3.1.1.59).
- Lipases (EC 3.1.1.3) from the fungi *Geotrichum candidum* and *Candida rugosa*.
- *Caenorhabditis* gut esterase (gene ges-1).
- Duck fatty acyl-CoA hydrolase, medium chain (EC 3.1.2.14), an enzyme that may be associated with peroxisome proliferation and may play a role in the production of 3-hydroxy fatty acid diester pheromones.
- Membrane enclosed crystal proteins from slime mold. These proteins are, most probably esterases; the vesicles where they are found have therefore

been termed esterosomes.

So far two bacterial proteins have been found to belong to this family:

- 5 - Phenmedipham hydrolase (phenylcarbamate hydrolase), an *Arthrobacter oxidans* plasmid-encoded enzyme (gene *pcd*) that degrades the phenylcarbamate herbicides phenmedipham and desmedipham by hydrolyzing their central carbamate linkages.
- Para-nitrobenzyl esterase from *Bacillus subtilis* (gene *pnbA*).

10

The following proteins, while having lost their catalytic activity, contain a domain evolutionary related to that of carboxylesterases type-B:

- 15 - Thyroglobulin (TG), a glycoprotein specific to the thyroid gland, which is the precursor of the iodinated thyroid hormones thyroxine (T4) and triiodo thyronine (T3).
- *Drosophila* protein neuractin (gene *nrt*) which may mediate or modulate cell adhesion between embryonic cells during development.
- *Drosophila* protein glutactin (gene *glt*), whose function is not known.

20

As is the case for lipases and serine proteases, the catalytic apparatus of esterases involves three residues (catalytic triad): a serine, a glutamate or aspartate and a histidine. The sequence around the active site serine is well conserved and can be used as a signature pattern. A conserved region located in
25 the N-terminal section containing a cysteine involved in a disulfide bond has been selected as a second signature pattern.

25

- Consensus pattern: F-[GR]-G-x(4)-[LIVM]-x-[LIV]-x-G-x-S-[STAG]-G[S is the active site residue]
- 30 -Consensus pattern: [ED]-D-C-L-[YT]-[LIV]-[DNS]-[LIV]-[LIVFYW]-x-[PQR] [C is involved in a disulfide bond]

30

[1] Myers M., Richmond R.C., Oakeshott J.G. Mol. Biol. Evol. 5:113-119(1988).

[2] Krejci E., Duval N., Chatonnet A., Vincens P., Massoulie J. Proc. Natl. Acad. Sci. U.S.A. 88:6647-6651(1991).

[3] Cygler M., Schrag J.D., Sussman J.L., Harel M., Silman I. Gentry M.K., Doctor B.P. Protein Sci. 2:366-382(1993).

5 [4] Lockridge O. BioEssays 9:125-128(1988).

[5] Wang C.-S., Hartsuck J.A. Biochim. Biophys. Acta 1166:1-19(1993).

85. CPSase_L_chain (Carbamoyl-phosphate synthase (CPSase))

10 [1]

Medline: 94347758

Three-dimensional structure of the biotin carboxylase subunit.
of acetyl-CoA carboxylase.

Waldrop GL, Rayment I, Holden HM;

Biochemistry 1994;33:10249-10256.

[1]

Medline: 90285162

Mammalian carbamyl phosphate synthetase (CPS). DNA sequence and
evolution of the CPS domain of the Syrian hamster multifunctional
protein CAD.

Simmer JP, Kelly RE, Rinker AG Jr, Scully JL, Evans DR;

Biol Chem 1990;265:10395-10402.

Carbamoyl-phosphate synthase catalyzes the ATP-dependent synthesis of
carbamyl-phosphate from glutamine or ammonia and bicarbonate. This
important enzyme initiates both the urea cycle and the biosynthesis
of arginine and/or pyrimidines [2].

The carbamoyl-phosphate synthase (CPS) enzyme in prokaryotes is a
heterodimer of a small and large chain. The small chain promotes
the hydrolysis of glutamine to ammonia, which is used by the large
chain to synthesize carbamoyl phosphate. See CPSase_sm_chain.

The small chain has a GATase domain in the carboxyl terminus.

See GATase.

Number of members: 181

Carbamoyl-phosphate synthase (CPSase) catalyzes the ATP-dependent synthesis of carbamyl-phosphate from glutamine (EC 6.3.5.5) or ammonia (EC 6.3.4.16) and bicarbonate [1]. This important enzyme initiates both the urea cycle and the biosynthesis of arginine and pyrimidines.

Glutamine-dependent CPSase (CPSase II) is involved in the biosynthesis of pyrimidines and purines. In bacteria such as *Escherichia coli*, a single enzyme is involved in both biosynthetic pathways while other bacteria have separate enzymes. The bacterial enzymes are formed of two subunits. A small chain (gene *carA*) that provides glutamine amidotransferase activity (GATase) necessary for removal of the ammonia group from glutamine, and a large chain (gene *carB*) that provides CPSase activity. Such a structure is also present in fungi for arginine biosynthesis (genes *CPA1* and *CPA2*). In most eukaryotes, the first three steps of pyrimidine biosynthesis are catalyzed by a large multifunctional enzyme - called *URA2* in yeast, rudimentary in *Drosophila* and *CAD* in mammals [2]. The CPSase domain is located between an N-terminal GATase domain and the C-terminal part which encompass the dihydroorotase and aspartate transcarbamylase activities.

Ammonia-dependent CPSase (CPSase I) is involved in the urea cycle in ureolytic vertebrates; it is a monofunctional protein located in the mitochondrial matrix.

The CPSase domain is typically 120 Kd in size and has arisen from the duplication of an ancestral subdomain of about 500 amino acids. Each subdomain independently binds to ATP and it is suggested that the two homologous halves act separately, one to catalyze the phosphorylation of bicarbonate to carboxy phosphate and the other that of carbamate to carbamyl phosphate.

The CPSase subdomain is also present in a single copy in the biotin-dependent enzymes acetyl-CoA carboxylase (EC 6.4.1.2) (ACC), propionyl-CoA carboxylase (EC 6.4.1.3) (PCCase), pyruvate carboxylase (EC 6.4.1.1) (PC) and urea

carboxylase (EC 6.3.4.6).

Two conserved regions which are probably important for binding ATP and/or catalytic activity have been selected as signatures for the subdomain.

5

-Consensus pattern: [FYV]-[PS]-[LIVMC]-[LIVMA]-[LIVM]-[KR]-[PSA]-[STA]-x(3)-[SG]-G-x-[AG]

-Consensus pattern: [LIVMF]-[LIMN]-E-[LIVMCA]-N-[PATLIVM]-[KR]-[LIVMSTAC]

10 [1] Simmer J.P., Kelly R.E., Rinker A.G. Jr., Scully J.L., Evans D.R.

J. Biol. Chem. 265:10395-10402(1990).

[2] Davidson J.N., Chen K.C., Jamison R.S., Musmanno L.A., Kern C.B.

BioEssays 15:157-164(1993).

15

86. CPSase_sm_chain (Carbamoyl-phosphate synthase small chain, CPSase domain)

[1]

Medline: 90285162

Mammalian carbamyl phosphate synthetase (CPS). DNA sequence and evolution of the CPS domain of the Syrian hamster multifunctional protein CAD.

20

Simmer JP, Kelly RE, Rinker AG Jr, Scully JL, Evans DR;

Biol Chem 1990;265:10395-10402.

The carbamoyl-phosphate synthase domain is in the amino terminus of protein.

25

Carbamoyl-phosphate synthase catalyzes the ATP-dependent synthesis of carbamyl-phosphate from glutamine or ammonia and bicarbonate. This important enzyme initiates both the urea cycle and the biosynthesis of arginine and/or pyrimidines [1].

30 The carbamoyl-phosphate synthase (CPS) enzyme in prokaryotes is a heterodimer of a small and large chain. The small chain promotes the hydrolysis of glutamine to ammonia, which is used by the large chain to synthesize carbamoyl phosphate. See CPSase_L_chain.

The small chain has a GATase domain in the carboxyl terminus.

See GATase.

Number of members: 46

5 Carbamoyl-phosphate synthase (CPSase) catalyzes the ATP-dependent synthesis of carbamyl-phosphate from glutamine (EC 6.3.5.5) or ammonia (EC 6.3.4.16) and bicarbonate [1]. This important enzyme initiates both the urea cycle and the biosynthesis of arginine and pyrimidines.

10 Glutamine-dependent CPSase (CPSase II) is involved in the biosynthesis of pyrimidines and purines. In bacteria such as Escherichia coli, a single enzyme is involved in both biosynthetic pathways while other bacteria have separate enzymes. The bacterial enzymes are formed of two subunits. A small chain (gene carA) that provides glutamine amidotransferase activity (GATase) necessary for
15 removal of the ammonia group from glutamine, and a large chain (gene carB) that provides CPSase activity. Such a structure is also present in fungi for arginine biosynthesis (genes CPA1 and CPA2). In most eukaryotes, the first three steps of pyrimidine biosynthesis are catalyzed by a large
20 multifunctional enzyme - called URA2 in yeast, rudimentary in Drosophila and CAD in mammals [2]. The CPSase domain is located between an N-terminal GATase domain and the C-terminal part which encompass the dihydroorotase and aspartate transcarbamylase activities.

25 Ammonia-dependent CPSase (CPSase I) is involved in the urea cycle in ureolytic vertebrates; it is a monofunctional protein located in the mitochondrial matrix.

30 The CPSase domain is typically 120 Kd in size and has arisen from the duplication of an ancestral subdomain of about 500 amino acids. Each subdomain independently binds to ATP and it is suggested that the two homologous halves act separately, one to catalyze the phosphorylation of bicarbonate to carboxy phosphate and the other that of carbamate to carbamyl phosphate.

The CPSase subdomain is also present in a single copy in the biotin-dependent enzymes acetyl-CoA carboxylase (EC 6.4.1.2) (ACC), propionyl-CoA carboxylase (EC 6.4.1.3) (PCCase), pyruvate carboxylase (EC 6.4.1.1) (PC) and urea carboxylase (EC 6.3.4.6).

5

Two conserved regions which are probably important for binding ATP and/or catalytic activity have been selected as signatures for the subdomain.

-Consensus pattern: [FYV]-[PS]-[LIVMC]-[LIVMA]-[LIVM]-[KR]-[PSA]-[STA]-x(3)-
[SG]-G-x-[AG]

10

-Consensus pattern: [LIVMF]-[LIMN]-E-[LIVMCA]-N-[PATLIVM]-[KR]-[LIVMSTAC]

[1] Simmer J.P., Kelly R.E., Rinker A.G. Jr., Scully J.L., Evans D.R.
J. Biol. Chem. 265:10395-10402(1990).

[2] Davidson J.N., Chen K.C., Jamison R.S., Musmanno L.A., Kern C.B.
BioEssays 15:157-164(1993).

87. CRAL_TRIO (CRAL/TRIO domain)

[1]

Medline: 98121119

Crystal structure of the *Saccharomyces cerevisiae* phosphatidyl-
inositol-transfer protein.

Sha B, Phillips SE, Bankaitis VA, Luo M;

Nature 1998;391:506-510.

The original profile has been extended to include the carboxyl
domain from the known structure of Sec14. Swiss:P10911 has not
been included in the Pfam family because it does not appear to
contain a complete structural domain.

Number of members: 39

88. CSD ('Cold-shock' DNA-binding domain)

[1]

Medline: 94255482

Crystal structure of CspA, the major cold shock protein of Escherichia coli.

- 5 Schindelin H, Jiang W, Inouye M, Heinemann U; Proc Natl Acad Sci U S A 1994;91:5119-5123.

Number of members: 121

A conserved domain of about 70 amino acids has been found in prokaryotic and eukaryotic DNA-binding proteins [1,2,3,E1]. This domain, which is known as the 'cold-shock domain' (CSD) is present in the proteins listed below.

- Escherichia coli protein CS7.4 (gene cspA) which is induced in response to low temperature (cold-shock protein) and which binds to and stimulates the transcription of the CCAAT-containing promoters of the HN-S protein and of gyrA.
- Mammalian Y box binding protein 1 (YB1). A protein that binds to the CCAAT-containing Y box of mammalian HLA class II genes.
- Xenopus Y box binding proteins -1 and -2 (Y1 and Y2). Proteins that bind to the CCAAT-containing Y box of Xenopus hsp70 genes.
- Xenopus B box binding protein (YB3). YB3 binds the B box promoter element of genes transcribed by RNA polymerase III.
- Enhancer factor I subunit A (EFI-A) (dbpB). A protein that also bind to CCAAT-motif in various gene promoters.
- DbpA, a Human DNA-binding protein of unknown specificity.
- Bacillus subtilis cold-shock proteins cspB and cspC.
- Streptomyces clavuligerus protein SC 7.0.
- Escherichia coli proteins cspB, cspC, cspD, cspE and cspF.
- Unr, a mammalian gene encoded upstream of the N-ras gene. Unr contains nine repeats that are similar to the CSD domain. The function of Unr is not yet known but it could be a multivalent DNA-binding protein.

As a signature pattern for the CSD domain, its most conserved

region which is located in its N-terminal section has been selected. It must be noted that the beginning of this region is highly similar [4] to the RNP-1 RNA-binding motif.

-Consensus pattern: [FY]-G-F-I-x(6,7)-[DER]-[LIVM]-F-x-H-x-[STKR]-x-[LIVMFY]

5

[1] Doniger J., Landsman D., Gonda M.A., Wistow G.
New Biol. 4:389-395(1992).

[2] Wistow G.
Nature 344:823-824(1990).

10 [3] Jones P.G., Inouye M.
Mol. Microbiol. 11:811-818(1994).

[4] Landsman D.
Nucleic Acids Res. 20:2861-2864(1992).

15

89. CTF_NFI (CTF/NF-I family)

Number of members: 45

20

Nuclear factor I (NF-I) or CCAAT box-binding transcription factor (CTF) [1,2] (also known as TGGCA-binding proteins) are a family of vertebrate nuclear proteins which recognize and bind, as dimers, the palindromic DNA sequence 5'-TGGCANNNTGCCA-3'. CTF/NF-I binding sites are present in viral and cellular promoters and in the origin of DNA replication of Adenovirus type 2.

25

The CTF/NF-I proteins were first identified as nuclear factor I, a collection of proteins that activate the replication of several Adenovirus serotypes (together with NF-II and NF-III) [3]. The family of proteins was also

30

identified as the CTF transcription factors, before the NFI and CTF families were found to be identical [4]. The CTF/NF-I proteins are individually capable of activating transcription and DNA replication. The CTF/NF-I family name has also been dubbed as NFI, NF-I or NF1.

In a given species, there are a large number of different CTF/NF-I proteins.

The multiplicity of CTF/NF-I is known to be generated both by alternative splicing and by the occurrence of four different genes. The known forms of NF-I genes have been classified as:

- 5 - The CTF-like factors subfamily (prototype form: CTF-1) [4]
- The NFI-X proteins.
- The NFI-A proteins.
- The NFI-B proteins.

10 So far, all CTF/NF-I family members appear to have similar transcription and replication activities.

CTF/NF-1 proteins contains 400 to 600 amino acids. The N-terminal 200 amino-
acid sequence, almost perfectly conserved in all species and genes sequenced,
15 mediates site-specific DNA recognition, protein dimerization and Adenovirus
DNA replication. The C-terminal 100 amino acids contain the transcriptional
activation domain. This activation domain is the target of gene expression
regulatory pathways elicited by growth factors and it interacts with basal
transcription factors and with histone H3 [6].

20 A perfectly conserved, highly charged 12 residue peptide located in the N-terminal part of
CTF/NF-I has been selected as a specific signature for this family of proteins.

-Consensus pattern: R-K-R-K-Y-F-K-K-H-E-K-R

25

[1] Mermod N., O'Neill E.A., Kelly T.J., Tjian R.
Cell 58:741-753(1989).

[2] Rupp R.A.W., Kruse U., Multhaup G., Goebel U., Beyreuther K.,
Sippel A.E.

30 Nucleic Acids Res. 18:2607-2616(1990).

[3] Nagata K., Guggenheimer R.A., Enomoto T., Lichy J.H., Hurwitz J.
Proc. Natl. Acad. Sci. U.S.A. 79:6438-6442(1982).

[4] Santoro C., Mermod N., Andrews P.C., Tjian R.

Nature 334:2118-2224(1988).

[5] Gil G., Smith J.R., Goldstein J.L., Slaughter C.A., Orth K., Brown M.S.,
Osborne T.F.

Proc. Natl. Acad. Sci. U.S.A 85:8963-8967(1988).

5 [6] Alevizopoulos A., Dusserre Y., Tsai-Pflugfelder M., von der Weid T.,
Wahli W., Mermod N.
Genes Dev. 9:3051-3066(1995).

10 90. Calsequestrin (Calsequestrin)

Number of members: 13

Calsequestrin is a moderate-affinity, high-capacity calcium-binding protein
of cardiac and skeletal muscle [1], where it is located in the luminal space
5 of the sarcoplasmic reticulum terminal cisternae. Calsequestrin acts as a
calcium buffer and plays an important role in the muscle excitation-
contraction coupling. It is a highly acidic protein of about 400 amino acid
residues that binds more than 40 moles of calcium per mole of protein. There
are at least two different forms of calsequestrin: one which is expressed in
20 cardiac muscles and another in skeletal muscles. Both forms have highly
similar sequences.

Two signature sequences have been developed. The first corresponds to the N-
terminus of the mature protein, the second is located just in front of the
25 C-terminus of the protein which is composed of a highly acidic tail of
variable length.

-Consensus pattern: [EQ]-[DE]-G-L-[DN]-F-P-x-Y-D-G-x-D-R-V

-Consensus pattern: [DE]-L-E-D-W-[LIVM]-E-D-V-L-x-G-x-[LIVM]-N-T-E-D-D-D

30

[1] Treves S., Vilsen B., Chiozzi P., Andersen J.P., Zorzato F.
Biochem. J. 283:767-772(1992).

91. Carboxyl_trans (Carboxyl transferase domain)

[1]

Medline: 93374821

5 Primary structure of the monomer of the 12S subunit of
transcarboxylase as deduced from DNA and characterization of the
product expressed in Escherichia coli.

Thornton CG, Kumar GK, Haase FC, Phillips NF, Woo SB, Park VM,
Magner WJ, Shenoy BC, Wood HG, Samols D;

10 J Bacteriol 1993;175:5301-5308.

[2]

Medline: 93358891

Molecular evolution of biotin-dependent carboxylases.

Toh H, Kondo H, Tanabe T;

5 Eur J Biochem 1993;215:687-696.

All of the members in this family are biotin dependent carboxylases.

The carboxyl transferase domain carries out the following reaction;
transcarboxylation from biotin to an acceptor molecule. There are
two recognised types of carboxyl transferase. One of them uses acyl-CoA
and the other uses 2-oxo acid as the acceptor molecule of carbon dioxide.

20 All of the members in this family utilise acyl-CoA as the acceptor
molecule.

Number of members: 47

25

92. Chal_stil_synt (Chalcone and stilbene synthases)

Number of members: 146

30

Chalcone synthases (CHS) (EC 2.3.1.74) and stilbene synthases (STS) (formerly
known as resveratrol synthases) are related plant enzymes [1]. CHS is an
important enzyme in flavanoid biosynthesis and STS a key enzyme in stilbene-
type phytoalexin biosynthesis. Both enzymes catalyze the addition of three
molecules of malonyl-CoA to a starter CoA ester (a typical example is

4-coumaroyl-CoA), producing either a chalcone (with CHS) or stilbene (with STS).

These enzymes are proteins of about 390 amino-acid residues. A conserved cysteine residue, located in the central section of these proteins, has been shown [2] to be essential for the catalytic activity of both enzymes and probably represents the binding site for the 4-coumaroyl-CoA group. The region around this active site residue is well conserved and can be used as a signature pattern.

In addition to the plant enzymes, this family also includes *Bacillus subtilis* bcsA.

-Consensus pattern: R-[LIVMFYS]-x-[LIVM]-x-[QHG]-x-G-C-[FYNA]-[GA]-G-[GA]-[STAV]-x-[LIVMF]-[RA] [C is the active site residue]

[1] Schroeder J., Schroeder G.

Z. Naturforsch. 45C:1-8(1990).

[2] Lanz T., Tropf S., Marner F.-J., Schroeder J., Schroeder G.

J. Biol. Chem. 266:9971-9976(1991).

93. Chorismate_synt (Chorismate synthase)

Number of members: 19

Chorismate synthase (EC 4.6.1.4) catalyzes the last of the seven steps in the shikimate pathway which is used in prokaryotes, fungi and plants for the biosynthesis of aromatic amino acids. It catalyzes the 1,4-trans elimination of the phosphate group from 5-enolpyruvylshikimate-3-phosphate (EPSP) to form chorismate which can then be used in phenylalanine, tyrosine or tryptophan biosynthesis. Chorismate synthase requires the presence of a reduced flavin mononucleotide (FMNH₂ or FADH₂) for its activity.

Chorismate synthase from various sources shows [1,2] a high degree of sequence conservation. It is a protein of about 360 to 400 amino-acid residues.

Three signature patterns have been developed from conserved regions rich in basic residues (mostly arginines). The first is in the N-terminal section, the second is central and the third is C-terminal.

-Consensus pattern: G-E-S-H-[GC]-x(2)-[LIVM]-[GTV]-x-[LIVM](2)-[DE]-G-x-[PV]

-Consensus pattern: [GE]-R-[SA](2)-[SAG]-R-[EV]-[ST]-x(2)-[RH]-V-x(2)-G

-Consensus pattern: R-[SH]-D-[PSV]-[CSAV]-x(4)-[GAI]-x-[IVGSP]-[LIVM]-x-E-[STAH]-[LIVM]

[1] Schaller A., Schmid J., Leibinger U., Amrhein N.
J. Biol. Chem. 266:21434-21438(1991).

[2] Jones D.G.L., Reusser U., Braus G.H.
Mol. Microbiol. 5:2143-2152(1991).

94. Clat_adaptor_s (Clathrin adaptor complex small chain)

Number of members: 21

Clathrin coated vesicles (CCV) mediate intracellular membrane traffic such as receptor mediated endocytosis. In addition to clathrin, the CCV are composed of a number of other components including oligomeric complexes which are known as adaptor or clathrin assembly proteins (AP) complexes [1]. The adaptor complexes are believed to interact with the cytoplasmic tails of membrane proteins, leading to their selection and concentration. In mammals two type of adaptor complexes are known: AP-1 which is associated with the Golgi complex and AP-2 which is associated with the plasma membrane. Both AP-1 and AP-2 are heterotetramers that consist of two large chains - the adaptins - (gamma and beta' in AP-1; alpha and beta in AP-2); a medium chain (AP47 in AP-1; AP50 in AP-2) and a small chain (AP19 in AP-1; AP17 in AP-2).

150

The small chains of AP-1 and AP-2 are evolutionary related proteins of about 18 Kd. Homologs of AP17 and AP19 have also been found in yeast (genes APS1/YAP19 and APS2/YAP17) [2,3,4]. AP17 and AP19 are also related to the zeta-chain [5] of coatamer (zeta-cop), a cytosolic protein complex that reversibly associates with Golgi membranes to form vesicles that mediate biosynthetic protein transport from the endoplasmic reticulum, via the Golgi up to the trans Golgi network.

A conserved region in the central section of these proteins has been selected as a signature pattern.

-Consensus pattern: [LIVM](2)-Y-[KR]-x(4)-L-Y-F

[1] Pearse B.M., Robinson M.S.

Annu. Rev. Cell Biol. 6:151-171(1990).

[2] Kirchhausen T., Davis A.C., Frucht S., O'Brine Greco B., Payne G.S., Tubb B.

J. Biol. Chem. 266:11153-11157(1991).

[3] Nakai M., Takada T., Endo T.

Biochim. Biophys. Acta 1174:282-284(1993).

[4] Phan H.L., Finlay J.A., Chu D.S., Tan P.K., Kirchhausen T., Payne G.S. EMBO J. 13:1706-1717(1994).

[5] Kuge O., Hara-Kuge S., Orci L., Ravazzola M., Amherdt M., Tanigawa G., Wieland F.T., Rothman J.E.

J. Cell Biol. 123:1727-1734(1993).

95. Clathrin_lg_ch (Clathrin light chain.)

Number of members: 8

Clathrin [1,2] is the major coat-forming protein that encloses vesicles such as coated pits and forms cell surface patches involved in membrane traffic within eukaryotic cells. The clathrin coats (called triskelions) are composed

of three heavy chains (180 Kd) and three light chains (23 to 27 Kd).

The clathrin light chains [3], which may help to properly orient the assembly and disassembly of the clathrin coats, bind non-covalently to the heavy chain, they also bind calcium and interact with the hsc70 uncoating ATPase.

- In higher eukaryotes two genes code for distinct but related light chains: LC(a) and LC(b). Each of the two genes can yield, by tissue-specific alternative splicing, two separate forms which differ by the insertion of a sequence of respectively thirty or eighteen residues. There is, in the N-terminal part of the clathrin light chains a domain of twenty one amino acid residues which is perfectly conserved in LC(a) and LC(b).
- In yeast there is a single light chain (gene CLC1) whose sequence is only distantly related to that of higher eukaryotes.

Two signature patterns have been developed for clathrin light chains. The first pattern is a heptapeptide from the center of the conserved N-terminal region of eukaryotic light chains; the second pattern is derived from a positively charged region located in the C-terminal extremity of all known clathrin light chains.

-Consensus pattern: F-L-A-Q-Q-E-S

[1] Keen J.H.

Annu. Rev. Biochem. 59:415-438(1990).

[2] Brodsky F.M.

Science 242:1396-1402(1988).

[3] Brodsky F.M., Hill B.L., Acton S.L., Naethke I., Wong D.H.,

Ponnambalam S., Parham P.

Trends Biochem. Sci. 16:208-213(1991).

96. (Clathrin repeat) 7-fold repeat in Clathrin and VPS

Each repeat is about 140 amino acids long. The repeats occur in the arm region of the Clathrin heavy chain.

Number of members: 79

[1]

5 Medline: 92191269

Folding and trimerization of clathrin subunits at the triskelion hub.

Nathke IS, Heuser J, Lupas A, Stock J, Turck CW, Brodsky FM;
Cell 1992;68:899-910. [2]

10 Medline: 88097376

Clathrin heavy chain: molecular cloning and complete primary structure.

Kirchhausen T, Harrison SC, Chow EP, Mattaliano RJ,
Ramachandran KL, Smart J, Brosius J;
15 Proc Natl Acad Sci U S A 1987;84:8805-8809.

97. Collagen (Collagen triple helix repeat (20 copies))

[1] Medline: 94059583

20 New members of the collagen superfamily

Mayne R, Brewton RG;
Curr Opin Cell Biol 1993;5:883-890.

Scurvy is associated with collagens.

Members of this family belong to the collagen superfamily [1].

25 Collagens are generally extracellular structural proteins involved in formation of connective tissue structure.

The alignment contains 20 copies of the G-X-Y repeat that forms a triple helix. The first position of the repeat is glycine, the second and third positions can be any residue

30 but are frequently proline and hydroxyproline. Collagens are post translationally modified by proline hydroxylase to form the hydroxyproline residues. Defective hydroxylation is the cause of scurvy.

Some members of the collagen superfamily are not involved in connective tissue structure but share the same triple helical structure.

Number of members: 2125

5

98. Coprogen_oxidas (Coproporphyrinogen III oxidase)

Number of members: 12

Coproporphyrinogen III oxidase (EC 1.3.3.3) (coproporphyrinogenase) [1,2] catalyzes the oxidative decarboxylation of coproporphyrinogen III into protoporphyrinogen IX, a common step in the pathway for the biosynthesis of porphyrins such as heme, chlorophyll or cobalamin.

10

Coproporphyrinogen III oxidase is an enzyme that requires iron for its activity. A cysteine seems to be important for the catalytic mechanism [3]. Sequences from a variety of eukaryotic and prokaryotic sources show that this enzyme has been evolutionarily conserved. A highly conserved region in the central part of the sequence has been selected as a signature pattern. This region contains the only conserved cysteine and is rich in charged amino acids.

15

20

-Consensus pattern: K-x-W-C-x(2)-[FYH](3)-[LIVM]-x-H-R-x-E-x-R-G-[LIVM]-G-G-[LIVM]-F-F-D

[1] Xu K., Elliott T.

J. Bacteriol. 175:4990-4999(1993).

25

[2] Kohno H., Furukawa T., Yoshinaga T., Tokunaga R., Taketani S.

J. Biol. Chem. 268:21359-21363(1993).

[3] Camadro J.M., Chambon H., Jolles J., Labbe P.

Eur. J. Biochem. 156:579-587(1986).

30

[4] Xu K., Elliott T.

J. Bacteriol. 176:3196-3203(1994).

99. Corona_nucleoca (Coronavirus nucleocapsid protein)

[1]

Medline: 98087828

5 Identification of a specific interaction between the coronavirus mouse hepatitis virus A59 nucleocapsid protein and packaging signal.

Molenkamp R, Spaan WJ;

Virology 1997;239:78-86.

10 Number of members: 44

100. Cu-oxidase (Multicopper oxidase)

[1]

15 Medline: 90126844

The blue oxidases, ascorbate oxidase, laccase and ceruloplasmin.

Modelling and structural relationships.

Messerschmidt A, Huber R;

Eur J Biochem 1990;187:341-352.

20 Number of members: 150

Multicopper oxidases [1,2] are enzymes that possess three spectroscopically different copper centers. These centers are called: type 1 (or blue), type 2 (or normal) and type 3 (or coupled binuclear). The enzymes that belong to this family are:

- Laccase (EC 1.10.3.2) (urishiol oxidase), an enzyme found in fungi and plants, which oxidizes many different types of phenols and diamines.

- Ascorbate oxidase (EC 1.10.3.3), a higher plant enzyme.

30 - Ceruloplasmin (EC 1.16.3.1) (ferroxidase), a protein found in the serum of mammals and birds, which oxidizes a great variety of inorganic and organic substances. Structurally ceruloplasmin exhibits internal sequence homology, and seem to have evolved from the triplication of a copper-binding domain

similar to that found in laccase and ascorbate oxidase.

In addition to the above enzymes there are a number of proteins which, on the basis of sequence similarities, can be said to belong to this family. These proteins are:

- Copper resistance protein A (copA) from a plasmid in *Pseudomonas syringae*. This protein seems to be involved in the resistance of the microbial host to copper.
- Blood coagulation factor V (Fa V).
- Blood coagulation factor VIII (Fa VIII) [E1].
- Yeast FET3 [3], which is required for ferrous iron uptake.
- Yeast hypothetical protein YFL041w and SpAC1F7.08, the fission yeast homolog.

Factors V and VIII act as cofactors in blood coagulation and are structurally similar [4]. Their sequence consists of a triplicated A domain, a B domain and a duplicated C domain; in the following order: A-A-B-A-C-C. The A-type domain is related to the multicopper oxidases.

Two signature patterns have been developed for these proteins. Both patterns are derived from the same region, which in ascorbate oxidase, laccase, in the third domain of ceruloplasmin, and in copA, contains five residues that are known to be involved in the binding of copper centers. The first pattern does not make any assumption on the presence of copper-binding residues and thus can detect domains that have lost the ability to bind copper (such as those in Fa V and Fa VIII), while the second pattern is specific to copper-binding domains.

- Consensus pattern: G-x-[FYW]-x-[LIVMFYW]-x-[CST]-x(8)-G-[LM]-x(3)-[LIVMFYW]
- Consensus pattern: H-C-H-x(3)-H-x(3)-[AG]-[LM]
- [The first two H's are copper type 3 binding residues]
- [The C, the 3rd H, and L or M are copper type 1 ligands]

101. Cullin (Cullin family)

Number of members: 24

5

The following proteins are collectively termed cullins [1]:

- *Caenorhabditis elegans* cul-1 (or lin-19), a protein required for developmentally programmed transitions from the G1 phase of the cell cycle to the G0 phase or the apoptotic pathway.
- *Caenorhabditis elegans* cul-2, cul-3, cul-4 (F45E12.3), cul-5 (ZK856.1) and cul-6 (K08E7.7).
- Mammalian CUL1, CUL2, CUL3, CUL4A and CUL4B.
- Mammalian vasopressin-activated calcium-mobilizing receptor (VACM-1), a kidney-specific protein thought to form a cell surface receptor [2] but which does not have any structural hallmarks of a receptor.
- *Drosophila* lin19.
- Yeast CDC53 [3], which acts in concert with CDC4 and UBC3 (CDC34) to control the G1-to-S phase transition.
- Yeast hypothetical protein YGR003w.
- Fission yeast hypothetical protein SpAC24H6.03.

The cullins are hydrophilic proteins of 740 to 815 amino acids. The C-terminal extremity is the most conserved part of these proteins. A signature pattern has been developed from that region.

-Consensus pattern: [LIV]-K-x(2)-[LIV]-x(2)-L-I-[DEQ]-[KRHNQ]-x-Y-[LIVM]-x-R-x(6,7)-[FY]-x-Y-x-[SA]>

[1] Kipreos E.T., Lander L.E., Wing J.P., He W.W., Hedgecock E.M.
Cell 85:829-839(1996).

[2] Burnatowska-Hledin M.A., Spielman W.S., Smith W.L., Shi P., Meyer J.M.,
Dewitt D.L.

Am. J. Physiol. 268:f1198-F1210(1995).

[3] Mathias N., Johnson S.L., Winey M., Adams A.E., Goetsch L., Pringle J.R.,
Byers B., Goebel M.G.

Mol. Cell. Biol. 16:6634-6643(1996).

5

102. (Cu_amine_oxid)

Copper amine oxidase signatures

Amine oxidases (AO) [1] are enzymes that catalyze the oxidation of a wide range of biogenic
amines including many neurotransmitters, histamine and xenobiotic amines. There are two
classes of amine oxidases: flavin-containing (EC 1.4.3.4) and copper-containing (EC 1.4.3.6).

Copper-containing AO is found in bacteria, fungi, plants and animals, it is an homodimeric
enzyme that binds one copper ion per subunit as well as a 2,4,5- trihydroxyphenylalanine
quinone (or topaquinone) (TPQ) cofactor. This cofactor is derived from a tyrosine residue.

Two signature patterns were derived for copper AO, the first one contains the tyrosine which
give rises to the TPQ cofactor while the second one contains one of the three histidines
that bind the copper atom [2].

Consensus pattern[LIVM]-[LIVMA]-[LIVMF]-x(4)-[ST]-x(2)-N-Y-[DE]-[YN] [The first Y
gives rises to TPQ] Sequences known to belong to this class detected by the patternALL.

Consensus patternT-x-[GS]-x(2)-H-[LIVMF]-x(3)-E-[DE]-x-P [H is a copper ligand]

Sequences known to belong to this class detected by the pattern ALL, except for lentil AO.

[1] Knowles P.F., Dooley D.M. (In) Metal ions in biological systems; Sigel H., Sigel A.,
Eds., 30:361- 403, Marcel Dekker, New-York, (1993).

[2] Parsons M.R., Convery M.A., Wilmot C.M., Yadav K.D.S., Blakeley V., Corner A.S.,
Phillips S.E.V., McPherson M.J., Knowles P.F. Structure 3:1171-1184(1995).

103. Cys-protease (Cysteine protease)

Number of members: 358

Eukaryotic thiol proteases (EC 3.4.22.-) [1] are a family of proteolytic enzymes which contain an active site cysteine. Catalysis proceeds through a thioester intermediate and is facilitated by a nearby histidine side chain; an asparagine completes the essential catalytic triad. The proteases which are currently known to belong to this family are listed below (references are only provided for recently determined sequences).

- Vertebrate lysosomal cathepsins B (EC 3.4.22.1), H (EC 3.4.22.16), L (EC 3.4.22.15), and S (EC 3.4.22.27) [2].
- Vertebrate lysosomal dipeptidyl peptidase I (EC 3.4.14.1) (also known as cathepsin C) [2].
- Vertebrate calpains (EC 3.4.22.17). Calpains are intracellular calcium-activated thiol protease that contain both a N-terminal catalytic domain and a C-terminal calcium-binding domain.
- Mammalian cathepsin K, which seems involved in osteoclastic bone resorption [3].
- Human cathepsin O [4].
- Bleomycin hydrolase. An enzyme that catalyzes the inactivation of the antitumor drug BLM (a glycopeptide).
- Plant enzymes: barley aleurain (EC 3.4.22.16), EP-B1/B4; kidney bean EP-C1, rice bean SH-EP; kiwi fruit actinidin (EC 3.4.22.14); papaya latex papain (EC 3.4.22.2), chymopapain (EC 3.4.22.6), caricain (EC 3.4.22.30), and proteinase IV (EC 3.4.22.25); pea turgor-responsive protein 15A; pineapple stem bromelain (EC 3.4.22.32); rape COT44; rice oryzain alpha, beta, and gamma; tomato low-temperature induced, Arabidopsis thaliana A494, RD19A and RD21A.
- House-dust mites allergens DerP1 and EurM1.
- Cathepsin B-like proteinases from the worms *Caenorhabditis elegans* (genes gcp-1, cpr-3, cpr-4, cpr-5 and cpr-6), *Schistosoma mansoni* (antigen SM31) and *Japonica* (antigen SJ31), *Haemonchus contortus* (genes AC-1 and AC-2), and *Ostertagia ostertagi* (CP-1 and CP-3).

- Slime mold cysteine proteinases CP1 and CP2.
- Cruzipain from *Trypanosoma cruzi* and *brucei*.
- Throphozoite cysteine proteinase (TCP) from various *Plasmodium* species.
- Proteases from *Leishmania mexicana*, *Theileria annulata* and *Theileria parva*.
- 5 - Baculoviruses cathepsin-like enzyme (v-cath).
- *Drosophila* small optic lobes protein (gene sol), a neuronal protein that contains a calpain-like domain.
- Yeast thiol protease BLH1/YCP1/LAP3.
- *Caenorhabditis elegans* hypothetical protein C06G4.2, a calpain-like
- 10 protein.

Two bacterial peptidases are also part of this family:

- Aminopeptidase C from *Lactococcus lactis* (gene pepC) [5].
- 15 - Thiol protease tpr from *Porphyromonas gingivalis*.

Three other proteins are structurally related to this family, but may have lost their proteolytic activity.

- 20 - Soybean oil body protein P34. This protein has its active site cysteine replaced by a glycine.
- Rat testin, a sertoli cell secretory protein highly similar to cathepsin L but with the active site cysteine is replaced by a serine. Rat testin should not be confused with mouse testin which is a LIM-domain protein (see
- 25 <PDOC00382>).
- *Plasmodium falciparum* serine-repeat protein (SERA), the major blood stage antigen. This protein of 111 Kd possesses a C-terminal thiol-protease-like domain [6], but the active site cysteine is replaced by a serine.

30 The sequences around the three active site residues are well conserved and can be used as signature patterns.

160

-Consensus pattern: Q-x(3)-[GE]-x-C-[YW]-x(2)-[STAGC]-[STAGCV] [C is the active site residue]

-Consensus pattern: [LIVMGSTAN]-x-H-[GSACE]-[LIVM]-x-[LIVMAT](2)-G-x-[GSADNH] [H is the active site residue]

5

-Consensus pattern: [FYCH]-[WI]-[LIVT]-x-[KRQAG]-N-[ST]-W-x(3)-[FYW]-G-x(2)-G-[LFYW]-[LIVMFYG]-x-[LIVMF] [N is the active site residue]

[1] Dufour E. Biochimie 70:1335-1342(1988).

10 [2] Kirschke H., Barrett A.J., Rawlings N.D. Protein Prof. 2:1587-1643(1995).

[3] Shi G.-P., Chapman H.A., Bhairi S.M., Deleeuw C., Reddy V.Y., Weiss S.J. FEBS Lett. 357:129-134(1995).

[4] Velasco G., Ferrando A.A., Puente X.S., Sanchez L.M., Lopez-Otin C. J. Biol. Chem. 269:27136-27142(1994).

15 [5] Chapot-Chartier M.P., Nardi M., Chopin M.C., Chopin A., Gripon J.C. Appl. Environ. Microbiol. 59:330-333(1993).

[6] Higgins D.G., McConnell D.J., Sharp P.M. Nature 340:604-604(1989).

[7] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:461-486(1994).

20

104. Cys_Met_Meta_PP (Cys/Met metabolism PLP-dependent enzyme)

[1] Medline: 96428687

Crystal structure of the pyridoxal-5'-phosphate dependent cystathionine beta-lyase from Escherichia coli at 1.83 Å.

25 Clausen T, Huber R, Laber B, Pohlenz HD, Messerschmidt A; J Mol Biol 1996;262:202-224.

[1] Medline: 99059720

Crystal structure of Escherichia coli cystathionine gamma-synthase at 1.5 Å resolution.

30 Clausen T, Huber R, Prade L, Wahl MC, Messerschmidt A; EMBO J 1998;17:6827-6838.

Database Reference: SCOP; 1cs1; fa; [SCOP-USA][CATH-PDBSUM]

This family includes enzymes involved in cysteine and

methionine metabolism. The following are members:

Cystathionine gamma-lyase,

Cystathionine gamma-synthase,

Cystathionine beta-lyase,

5 Methionine gamma-lyase,

OAH/OAS sulfhydrylase,

O-succinylhomoserine sulphhydrylase

All of these members participate in slightly different reactions.

All these enzymes use PLP (pyridoxal-5'-phosphate) as a cofactor.

10 Number of members: 52

A number of pyridoxal-dependent enzymes involved in the metabolism of cysteine, homocysteine and methionine have been shown [1,2] to be evolutionary related. These are:

- 15 - Cystathionine gamma-lyase (EC 4.4.1.1) (gamma-cystathionase), which catalyzes the transformation of cystathionine into cysteine, oxobutanoate and ammonia. This is the final reaction in the transsulfuration pathway that leads from methionine to cysteine in eukaryotes.
- 20 - Cystathionine gamma-synthase (EC 4.2.99.9), which catalyzes the conversion of cysteine and succinyl-homoserine into cystathionine and succinate: the first step in the biosynthesis of methionine from cysteine in bacteria (gene metB).
- 25 - Cystathionine beta-lyase (EC 4.4.1.8) (beta-cystathionase), which catalyzes the conversion of cystathionine into homocysteine, pyruvate and ammonia: the second step in the biosynthesis of methionine from cysteine in bacteria (gene metC).
- 30 - Methionine gamma-lyase (EC 4.4.1.11) (L-methioninase) which catalyzes the transformation of methionine into methanethiol, oxobutanoate and ammonia.
- OAH/OAS sulfhydrylase, which catalyzes the conversion of acetylhomoserine into homocysteine and that of acetylserine into cysteine (gene MET17 or MET25 in yeast).
- O-succinylhomoserine sulfhydrylase (EC 4.2.99.-).

- Yeast hypothetical protein YGL184c.

- Yeast hypothetical protein YHR112c.

These enzymes are proteins of about 400 amino-acid residues. The pyridoxal-P
 5 group is attached to a lysine residue located in the central section of these
 enzymes; the sequence around this residue is highly conserved and can be used
 as a signature pattern to detect this class of enzymes.

-Consensus pattern: [DQ]-[LIVMF]-x(3)-[STAGC]-[STAGCI]-T-K-[FYWQ]-[LIVMF]-x-G-
 10 [HQ]-[SGNH] [K is the pyridoxal-P attachment site]

[1] Ono B.I., Tanaka K., Naito K., Heike C., Shinoda S., Yamamoto S.,
 Ohmori S., Oshima T., Toh-E A.
 J. Bacteriol. 174:3339-3347(1992).

15 [2] Barton A.B., Kaback D.B., Clark M.W., Keng T., Ouellette B.F.F.,
 Storms R.K., Zeng B., Zhong W.W., Fortin N., Delaney S., Bussey H.
 Yeast 9:363-369(1993).

20 105. Cyt₂ reductase

FAD/NAD-binding Cytochrome reductase

Number of members: 60

[1] Medline: 95111952

Crystal structure of the FAD-containing fragment of corn
 25 nitrate reductase at 2.5 Å resolution: relationship to other
 flavoprotein reductases.

Lu G, Campbell WH, Schneider G, Lindqvist Y;
 Structure 1994;2:809-821.

[2] Medline: 92084635

30 The sequence of squash NADH:nitrate reductase and its
 relationship to the sequences of other flavoprotein
 oxidoreductases. A family of flavoprotein pyridine
 nucleotide cytochrome reductases.

Hyde GE, Crawford NM, Campbell WH;
J Biol Chem 1991;266:23542-23547.

5 106. Cytidylyltrans

Phosphatidate cytidylyltransferase

Number of members: 21

10 Phosphatidate cytidylyltransferase (EC 2.7.7.41) [1,2,3] (also known as CDP-diacylglycerol synthase) (CDS) is the enzyme that catalyzes the synthesis of CDP-diacylglycerol from CTP and phosphatidate (PA). CDP-diacylglycerol is an important branch point intermediate in both prokaryotic and eukaryotic organisms. CDS is a membrane-bound enzyme. A conserved region located in the C-terminal part has been selected as a signature pattern.

15 -Consensus pattern: S-x-[LIVMF]-K-R-x(4)-K-D-x-[GSA]-x(2)-[LI]-[PG]-x-H-G-G-[LIVM]-x-D-R-[LIVMF]-D

[1] Sparrow C.P., Raetz C.R.H.

J. Biol. Chem. 260:12084-12091(1985).

20 [2] Shen H., Heacock P.N., Clancey C.J., Dowhan W.

J. Biol. Chem. 271:789-795(1996).

[3] Saito S., Goto K., Tonosaki A., Kondo H.

J. Biol. Chem. 272:9503-9509(1997).

25

107. (Cytidylyltransf) Cytidylyltransferase. This family includes: Cholinephosphate cytidylyltransferase. Glycerol-3-phosphate cytidylyltransferase.

Number of members: 64

30

[1] Medline: 10208837 CTP:Phosphocholine Cytidylyltransferase: Insights into Regulatory Mechanisms and Novel Functions. Clement JM, Kent C; Biochem Biophys Res Commun 1999;257:643-650.

108. (cNMP binding) Cyclic nucleotide-binding domain signatures and profile

Proteins that bind cyclic nucleotides (cAMP or cGMP) share a structural domain of about 120

residues [1-3]. The best studied of these proteins is the prokaryotic catabolite gene activator (also known as the cAMP receptor protein) (gene *crp*) where such a domain is known to be composed of three α -helices and a distinctive eight-stranded, antiparallel β -barrel structure. Such a domain is known to exist in the following proteins: - Prokaryotic

catabolite gene activator protein (CAP). - cAMP- and cGMP-dependent protein kinases (cAPK and cGPK). Both types of kinases contain two tandem copies of the cyclic

nucleotide-binding domain. The cAPK's are composed of two different subunits: a catalytic chain and a regulatory chain which contains both copies of the domain. The cGPK's are single chain enzymes that include the two copies of the domain in their N- terminal section. The nucleotide specificity of cAPK and cGPK is due to an amino acid in the conserved

region of β -barrel 7: a threonine that is invariant in cGPK is an alanine in most cAPK. - Vertebrate cyclic nucleotide-gated ion-channels. Two such cation channels have been fully characterized. One is found in rod cells where it plays a role in visual signal transduction. It specifically binds to cGMP leading to an opening of the channel and thereby causing a depolarization of rod photoreceptors. In olfactory epithelium a similar, cAMP-binding, channel plays a role in odorant signal transduction. There are six invariant amino acids in this domain, three of which are glycine residues that are thought to be essential for

maintenance of the structure of the β -barrel. Two signature patterns for this domain have been developed. The first pattern is located within β -barrels 2 and 3 and contains the first two conserved Gly. The second pattern is located within β -barrels 6 and 7 and contains the third conserved Gly as well as the three other invariant residues.-

First consensus pattern: [LIVM]-[VIC]-x(2)-G-[DENQTA]-x-[GAC]-x(2)-[LIVMFY](4)-x(2)-G

Second consensus pattern: [LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV]-

[1] Weber I.T., Shabb J.B., Corbin J.D. Biochemistry 28:6122-6127(1989).

[2] Kaupp U.B. Trends Neurosci. 14:150-157(1991).

[3] Shabb J.B., Corbin J.D. J. Biol. Chem. 267:5723-5726(1992).

109. (cadherin)

Cadherins extracellular repeated domain signature

5 Cadherins [1,2] are a family of animal glycoproteins responsible for calcium-dependent cell-cell adhesion. Cadherins preferentially interact with themselves in a homophilic manner in connecting cells; thus acting as both receptor and ligand. A wide number of tissue-specific forms of cadherins are known:

- 10 - Epithelial (E-cadherin) (also known as uvomorulin or L-CAM) (CDH1).
- Neural (N-cadherin) (CDH2).
- Placental (P-cadherin) (CDH3).
- Retinal (R-cadherin) (CDH4).
- Vascular endothelial (VE-cadherin) (CDH5).
15 - Kidney (K-cadherin) (CDH6).
- Cadherin-8 (CDH8).
- Osteoblast (OB-cadherin) (CDH11).
- Brain (BR-cadherin) (CDH12).
- T-cadherin (truncated cadherin) (CDH13).
20 - Muscle (M-cadherin) (CDH14).
- Liver-intestine (LI-cadherin).
- EP-cadherin.

Structurally, cadherins are built of the following domains: a signal sequence, followed by a
25 propeptide of about 130 residues, then an extracellular domain of around 600 residues, then a transmembrane region, and finally a C-terminal cytoplasmic domain of about 150 residues. The extracellular domain can be sub- divided into five parts: there are four repeats of about 110 residues followed by a region that contains four conserved cysteines. It is suggested that the calcium-binding region of cadherins is located in the extracellular repeats.

30 Cadherins are evolutionary related to the desmogleins which are component of intercellular desmosome junctions involved in the interaction of plaque proteins:

- Desmoglein 1 (desmosomal glycoprotein I).
- Desmoglein 2.
- Desmoglein 3 (Pemphigus vulgaris antigen).

- 5 The Drosophila fat protein [3] is a huge protein of over 5000 amino acids that contains 34 cadherin-like repeats in its extracellular domain.

10 The signature pattern that was developed for the repeated domain is located in it the C-terminal extremity which is its best conserved region. The pattern includes two conserved aspartic acid residues as well as two asparagines; these residues could be implicated in the binding of calcium.

15 Consensus pattern[LIV]-x-[LIV]-x-D-x-N-D-[NH]-x-P Sequences known to belong to this class detected by the pattern ALL. Note this pattern is found in the first, second, and fourth copies of the repeated domain. In the third copy there is a deletion of one residue after the second conserved Asp.

- 20 [1] Takeichi M. Annu. Rev. Biochem. 59:237-252(1990).
 [2] Takeichi M. Trends Genet. 3:213-217(1987).
 [3] Mahoney P.A., Weber U., Onofrechuk P., Biessmann H., Bryant P.J., Goodman C.S. Cell 67:853-868(1991).

110. Calreticulin family signatures

- 25 Calreticulin [1] (also known as calregulin, CRP55 or HACBP) is a high-capacitycalcium-binding protein which is present in most tissues and located at the periphery of the endoplasmic (ER) and the sarcoplasmic reticulum (SR)membranes. It probably plays a role in the storage of calcium in the lumen ofthe ER and SR and it may well have other important functions. Structurally, calreticulin is a protein of about 400 amino acid residues consisting of
 30 three domains: a) An N-terminal, probably globular, domain of about 180 amino acid residues (N-domain); b) A central domain of about 70 residues (P-domain) which contains three repeats of an acidic 17 amino acid motif. This region binds calcium with a low-capacity, but a high-affinity; c) A C-terminal domain rich in acidic residues and in lysine (C-

domain). This region binds calcium with a high-capacity but a low-affinity. Calreticulin is evolutionary related to the following proteins: - *Onchocerca volvulus* antigen RAL-1. RAL-1 is highly similar to calreticulin, but possesses a C-terminal domain rich in lysine and arginine and lacks acidic residues and is therefore not expected to bind calcium in that region. -

5 Calnexin [2]. A calcium-binding protein that interacts with newly synthesized glycoproteins in the endoplasmic reticulum. It seems to play a major role in the quality control apparatus of the ER by the retention of incorrectly folded proteins. - Calmegin [3] (or calnexin-T), a testis-specific calcium-binding protein highly similar to calnexin. Three signature patterns have been developed for this family of proteins. The first two patterns are based on conserved
10 regions in the N-domain; the third pattern corresponds to positions 4 to 16 of the repeated motif in the P-domain.

Consensus pattern: [KRHN]-x-[DEQN]-[DEQNK]-x(3)-C-G-G-[AG]-[FY]-[LIVM]-[KN]-[LIVMFY](2)-

Consensus pattern: [LIVM](2)-F-G-P-D-x-C-[AG]-

15 Consensus pattern: [IV]-x-D-x-[DENST]-x(2)-K-P-[DEH]-D-W-[DEN]-

[1] Michalak M., Milner R.E., Burns K., Opas M. *Biochem. J.* 285:681-692(1992).

[2] Bergeron J.J.M., Brenner M.B., Thomas D.Y., Williams D.B. *Trends Biochem. Sci.* 19:124-128(1994).

20 [3] Watanabe D., Yamada K., Nishina Y., Tajima Y., Koshimizu U., Nagata A., Nishimune Y. *J. Biol. Chem.* 269:7744-7749(1994).

111. Eukaryotic-type carbonic anhydrases signature (carb_anhydrase)

25 Carbonic anhydrases (EC 4.2.1.1) (CA) [1,2,3,4] are zinc metalloenzymes which catalyze the reversible hydration of carbon dioxide. Eight enzymatic and evolutionary related forms of carbonic anhydrase are currently known to exist in vertebrates: three cytosolic isozymes (CA-I, CA-II and CA-III); two membrane-bound forms (CA-IV and CA-VII); a mitochondrial form (CA-V); a secreted salivary form (CA-VI); and a yet uncharacterized isozyme [5]. In the
30 alga *Chlamydomonas reinhardtii*, two CA isozymes have been sequenced[6]. They are periplasmic glycoproteins evolutionary related to vertebrate CAs. Some bacteria, such as *Neisseria gonorrhoeae* [7] also have a eukaryotic-type CA. CAs contain a single zinc atom bound to three conserved histidine residues. As a signature for CAs, a pattern has been

developed which includes one of these zinc-binding histidines. Protein D8 from Vaccinia and other poxviruses is related to CAs but has lost two of the zinc-binding histidines as well as many otherwise conserved residues. This is also true of the N-terminal extracellular domain of some receptor-type tyrosine-protein phosphatases (see <[PDOC00323](#)>).

- 5 Consensus pattern: S-E-[HN]-x-[LIVM]-x(4)-[FYH]-x(2)-E-[LIVMGA]-H-[LIVMFA](2)
[The second H is a zinc ligand]-

Note: most prokaryotic CA's as well as plant chloroplast CA's belong to another, evolutionary distinct family of proteins (see <[PDOC00586](#)

- 10 [1] Deutsch H.F. Int. J. Biochem. 19:101-113(1987).
[2] Fernley R.T. Trends Biochem. Sci. 13:356-359(1988).
[3] Tashian R.E. BioEssays 10:186-192(1989).
[4] Edwards Y. Biochem. Soc. Trans. 18:171-175(1990).
[5] Skaggs L.A., Bergenheim N.C.H., Venta P.J., Tashian R.E. Gene 126:291-292(1993).
15 [6] Fujiwara S., Fukuzawa H., Tachiki A., Miyachi S. Proc. Natl. Acad. Sci. U.S.A. 87:9779-9783(1990).
[7] Huang S., Xue Y., Sauer-Eriksson E., Chirica L., Lindskog S., Jonsson B.H. 2.3.CO;2-J. Mol. Biol. 283:301-310(1998).

20 112. Caseins alpha/beta signature

Caseins [1] are the major protein constituent of milk. Caseins can be classified into two families; the first consists of the kappa-caseins, and the second groups the alpha-s1, alpha-s2, and beta-caseins. The alpha/beta caseins are a rapidly diverging family of proteins. However
25 two regions are conserved: a cluster of phosphorylated serine residues and the signal sequence. The signature pattern has been developed for this family of proteins based upon the last eight residues of the signal sequence.

Consensus pattern: C-L-[LV]-A-x-A-[LVF]-A –

- 30 [1] Holt C., Sawyer L. Protein Eng. 2:251-259(1988).

113. Catalase signatures

Catalase (EC 1.11.1.6) [1,2,3] is an enzyme, present in all aerobic cells, that decomposes hydrogen peroxide to molecular oxygen and water. Its main function is to protect cells from the toxic effects of hydrogen peroxide. In eukaryotic organisms and in some prokaryotes catalase is a molecule composed of four identical subunits. Each of the subunits binds one

5 protoheme IX group. A conserved tyrosine serves as the heme proximal side ligand. The region around this residue has been used as a first signature pattern; it also includes a conserved arginine that participates in heme-binding. A conserved histidine has been shown to be important for the catalytic mechanism of the enzyme. The region around this residue has been selected as a second signature pattern.-

10 Consensus pattern: R-[LIVMFSTAN]-F-[GASTNP]-Y-x-D-[AST]-[QEH] [Y is the proximal heme-binding ligand]

Consensus pattern: [IF]-x-[RH]-x(4)-[EQ]-R-x(2)-H-x(2)-[GAS]-[GASTF]-[GAST] [H is an active site residue]

Note: some prokaryotic catalases belong to the peroxidase family (see <PDOC00394>).

[1] Murthy M.R.N., Reid T.J. III, Sicignano A., Tanaka N., Rossmann M.G. J. Mol. Biol. 152:465-499(1981).

[2] Melik-Adamyan W.R., Barynin V.V., Vagin A.A., Borisov V.V., Vainshtein B.K., Fita I., Murthy M.R.N., Rossmann M.G. J. Mol. Biol. 188:63-72(1986).

[3] von Ossowski I., Hausner G., Loewen P.C. J. Mol. Evol. 37:71-76(1993).

114. (chitin binding) Chitin recognition or binding domain signature

A conserved domain of 43 amino acids is found in several plant and fungal proteins that have a common binding specificity for oligosaccharides of N-acetylglucosamine [1]. This domain may be involved in the recognition or binding of chitin subunits. It has been found in the proteins listed below. - A number of non-leguminous plant lectins. The best characterized of these lectins are the three highly homologous wheat germ agglutinins (WGA-1, 2 and 3). WGA is an N-acetylglucosamine/N-acetylneuraminic acid binding lectin which structurally

25 consists of a fourfold repetition of the 43 amino acid domain. The same type of structure is found in a barley root-specific lectin as well as a rice lectin. - Plants endochitinases (EC 3.2.1.14) from class IA (see <PDOC00620>). Endochitinases are enzymes that catalyze the hydrolysis of the beta-1,4 linkages of N-acetyl glucosamine polymers of chitin. Plant

30

chitinases function as a defense against chitin containing fungal pathogens. Class IA chitinases generally contain one copy of the chitin-binding domain at their N-terminal extremity. An exception is agglutinin/chitinase [2] from the stinging nettle *Urtica dioica* which contains two copies of the domain. - Hevein [5], a wound-induced protein found in the latex of rubber trees. - Win1 and win2, two wound-induced proteins from potato. -

Kluyveromyces lactis killer toxin alpha subunit [3]. The toxin encoded by the linear plasmid pGKL1 is composed of three subunits: alpha, beta, and gamma. The gamma subunit harbors toxin activity and inhibits growth of sensitive yeast strains in the G1 phase of the cell cycle; the alpha subunit, which is proteolytically processed from a larger precursor that also contains the beta subunit, is a chitinase (see <PDOC00839>). In chitinases, as well as in the potato wound-induced proteins, the 43-residue domain directly follows the signal sequence and is therefore at the N-terminal of the mature protein; in the killer toxin alpha subunit it is located in the central section of the protein. The domain contains eight conserved cysteine residues which have all been shown, in WGA, to be involved in disulfide bonds. The topological arrangement of the four disulfide bonds is shown in the following figure: +-----
 -----+ +----|-----+ |||| xxCgxxxxxxxxCxxxxCCsxxgxCgxxxxxCxxxCxxxxC |
 *****|***** |||| +----+ +-----+'C': conserved cysteine involved in a
 disulfide bond. '*' : position of the pattern.

-Consensus pattern: C-x(4,5)-C-C-S-x(2)-G-x-C-G-x(4)-[FYW]-C [The five C's are involved in disulfide bonds]

[1] Wright H.T., Sandrasegaram G., Wright C.S. J. Mol. Evol. 33:283-294(1991).

[2] Lerner D.R., Raikhel N.V. J. Biol. Chem. 267:11085-11091(1992).

[3] Butler A.R., O'Donnel R.W., Martin V.J., Gooday G.W., Stark M.J.R. Eur. J. Biochem. 199:483-488(1991).

115. (Chitinase 1) Chitinases family 19 signatures

Chitinases (EC 3.2.1.14) [1] are enzymes that catalyze the hydrolysis of the beta-1,4-N-acetyl-D-glucosamine linkages in chitin polymers. From the viewpoint of sequence similarity chitinases belong to either family 18 or 19 in the classification of glycosyl hydrolases [2,E1]. Chitinases of family 19(also known as classes IA or I and IB or II) are enzymes from plants

171

that function in the defense against fungal and insect pathogens by destroying their chitin-containing cell wall. Class IA/I and IB/II enzymes differ in the presence (IA/I) or absence (IB/II) of a N-terminal chitin-binding domain (see the relevant entry <PDOC00025>). The catalytic domain of these enzymes consist of about 220 to 230 amino acid residues. Two highly conserved regions have been selected as signature patterns, the first one is located in the N-terminal section and contains one of the six cysteines which are conserved in most, if not all, of these chitinases and which is probably involved in a disulfide bond.

Consensus pattern: C-x(4,5)-F-Y-[ST]-x(3)-[FY]-[LIVMF]-x-A-x(3)-[YF]-x(2)-F-[GSA]

Consensus pattern: [LIVM]-[GSA]-F-x-[STAG](2)-[LIVMFY]-W-[FY]-W-[LIVM]

[1] Flach J., Pilet P.-E., Jolles P. Experientia 48:701-716(1992).

[2] Henrissat B. Biochem. J. 280:309-316(1991).

116. chloroa_b-bind

Chlorophyll A-B binding proteins. Number of members: 211

117. chromo

The 'chromo' (CHRomatin Organization MOdifier) domain [1 to 4] is a conserved region of about 60 amino acids which was originally found in Drosophila modifiers of variegation, which are proteins that modify the structure of chromatin to the condensed morphology of heterochromatin, a cytologically visible condition where gene expression is repressed. In protein Polycomb, the chromo domain has been shown to be important for chromatin targeting. Proteins that contains a chromo domain seem to fall into three classes:

- a) Proteins which have a N-terminal chromo domain followed by a region which is related to but distinct from the chromo domain and which has been termed [3] the 'chromo shadow' domain.
- b) Proteins with a single chromo domain.
- c) Proteins with paired tandem chromo domains.

Currently, this domain has been found in the following proteins:

Class A.

- Drosophila heterochromatin protein Su(var)205 (HP1).
- Human heterochromatin protein HP1 alpha.
- Mammalian modifier 1 and modifier 2.
- Fission yeast swi6, a protein involved in the repression of the silent mating-type loci mat2 and mat3.

Class B.

- Drosophila protein Polycomb (Pc).
- Mammalian modifier 3, a homolog of Pc.
- Drosophila protein Su(var)3-9, a suppressor of position-effect variegation.
- Human Mi-2 autoantigen, characteristic of dermatomyositis.
- Fungal retrotransposon polyproteins: 'skippy' from *Fusarium oxysporum*, 'grasshopper' and 'MAGGY' from *Magnaporthe grisea* and Cft-1 from *Cladosporium fulvum*.
- Fission yeast hypothetical protein SpAC18G6.02c.
- *Caenorhabditis elegans* hypothetical protein C29H12.5
- *Caenorhabditis elegans* hypothetical protein ZK1236.2.
- *Caenorhabditis elegans* hypothetical protein T09A5.8.

Class C.

- Mammalian DNA-binding/helicase proteins CHD-1 to CHD-4.
- Yeast protein CHD1.

The signature pattern for this domain corresponds to its best conserved section, which is located in its central part.

-Consensus pattern: [FYL]-x-[LIVMC]-[KR]-W-x-[GDNR]-[FYWLME]-x(5,6)-[ST]-W-[ESV]-[PSTDEN]-x(2,3)-[LIVMC]

. 173

[1] Paro R. Trends Genet. 6:416-421(1990).

[2] Singh P.B., Miller J.R., Pearce J., Kothary R., Burton R.D., Paro R., James T.C., Gaunt S.J. Nucleic Acids Res. 19:789-794(1991).

[3] Aasland R., Stewart A.F. Nucleic Acids Res. 23:3168-3173(1995).

5 [4] Koonin E.V., Zhou S., Lucchesi J.C. Nucleic Acids Res. 23:4229-4233(1995).

118. citrate_synt

10 Citrate synthase (EC 4.1.3.7) (CS) is the tricarboxylic acid cycle enzyme that catalyzes the synthesis of citrate from oxaloacetate and acetyl-CoA in an aldol condensation. CS can directly form a carbon-carbon bond in the absence of metal ion cofactors.

5 In prokaryotes, citrate synthase is composed of six identical subunits. In eukaryotes, there are two isozymes of citrate synthase: one is found in the mitochondrial matrix, the second is cytoplasmic. Both seem to be dimers of identical chains.

20 There are a number of regions of sequence similarity between prokaryotic and eukaryotic citrate synthases. One of the best conserved contains a histidine which is one of three residues shown [1] to be involved in the catalytic mechanism of the vertebrate mitochondrial enzyme. This region has been used as a signature pattern.

25 -Consensus pattern: G-[FYA]-[GA]-H-x-[IV]-x(1,2)-[RKT]-x(2)-D-[PS]-R [H is an active site residue]

[1] Karpusas M., Branchaud B., Remington S.J. Biochemistry 29:2213-2219(1990).

30

119. clpA_B

Chaperonin clpA/B

CAUTION! This family is a subfamily of the AAA

superfamily. The threshold has been set very high to stop overlaps with the AAA superfamily. This entry will be subsumed by AAA in the future.

Number of members: 39

5

A number of ATP-binding proteins that are thought to protect cells from extreme stress by controlling the aggregation of denaturation of vital cellular structures have been shown [1,2] to be evolutionary related. These proteins are listed below.

10

- Escherichia coli clpA, which acts as the regulatory subunit of the ATP-dependent protease clp.
- Rhodopseudomonas blastica clpA homolog.
- Escherichia coli heat shock protein clpB and homologs in other bacteria.
- Bacillus subtilis protein mecB.
- Yeast heat shock protein 104 (gene HSP104), which is vital for tolerance to heat, ethanol and other stresses.
- Neurospora heat shock protein hsp98.
- Yeast mitochondrial heat shock protein 78 (gene HSP78) [3].
- CD4A and CD4b, two highly related tomato proteins that seem to be located in the chloroplast.
- Trypanosoma brucei protein clp.
- Porphyra purpurea chloroplast encoded clpC.

25

The size of these proteins range from 84 Kd (clpA) to slightly more than 100 Kd (HSP104). They all share two conserved regions of about 200 amino acids that each contains an ATP-binding site. In addition to the ATP-binding A and B motifs there are many parts in these two domains that are also conserved. Two of these regions have been selected as signature patterns. The first signature is located in the first domain, some ten residues to the C-terminal of the ATP-binding B motif. The second pattern is located in the second domain in-between the ATP-binding A and B motifs.

30

-Consensus pattern: D-[AI]-[SGA]-N-[LIVMF](2)-K-[PT]-x-L-x(2)-G

-Consensus pattern: R-[LIVMFY]-D-x-S-E-[LIVMFY]-x-E-[KRQ]-x-[STA]-x-[STA]-[KR]-
[LIVM]-x-G-[STA]

- 5 [1] Gottesman S., Squires C., Pichersky E., Carrington M., Hobbs M., Mattick J.S.,
Dalrymple B., Kuramitsu H., Shiroza T., Foster T., Clark W.P., Ross B., Squires C.L.,
Maurizi M.R. Proc. Natl. Acad. Sci. U.S.A. 87:3513-3517(1990).
[2] Parsell D.A., Sanchez Y., Stitzel J.D., Lindquist S. Nature 353:270-273(1991).
[3] Leonhardt S.A., Fearon K., Danese P.N., Mason T.L. Mol. Cell. Biol. 13:6304-
10 6313(1993).

120. cofilin_ADF

Cofilin/tropomyosin-type actin-binding proteins

[1]

Medline: 97290449

Structure determination of yeast cofilin.

Fedorov AA, Lappalainen P, Fedorov EV, Drubin DG, Almo SC;
Nat Struct Biol 1997;4:366-369.

[2]

Medline: 97290450

Crystal structure of the actin-binding protein actophorin
from Acanthamoeba.

Leonard SA, Gittis AG, Petrella EC, Pollard TD, Lattman EE;
25 Nat Struct Biol 1997;4:369-373.

[3]

Medline: 97420794

F-actin and G-actin binding are uncoupled by mutation of
conserved tyrosine residues in maize actin depolymerizing
30 factor.

Jiang CJ, Weeds AG, Khan S, Hussey PJ;
Proc Natl Acad Sci U S A 1997;94:9973-9978.

[4]

Medline: 97357155

Cofilin promotes rapid actin filament turnover in vivo.

Lappalainen P, Drubin DG;

Nature 1997;388:78-82.

5 Severs actin filaments and binds to actin monomers.

Number of members: 44

10 Actin-depolymerizing proteins sever actin filaments (F-actin) and/or bind to actin monomers, or G-actin, thus preventing actin-polymerization by sequestering the monomers. The following proteins are evolutionary related and belong to a family of low molecular weight (137 to 166 residues) actin-depolymerizing proteins [1,2,3,4]:

- 15 - Cofilin from vertebrates, slime mold and yeast. Cofilin binds to F-actin and acts as a pH-dependent actin-depolymerizing protein.
- Destrin from vertebrates. Destrin binds to G-actin in a pH-independent manner and prevents polymerization.
- Caenorhabditis elegans unc-60.
- Acanthamoeba castellanii actophorin.
- 20 - Plants actin depolymerizing factor (ADF).

The most conserved region of these proteins is a twenty amino-acid segment that ends some 30 residues from their C-terminal extremity. This segment has been shown [5] to be important for actin-binding.

25 -Consensus pattern: P-[DE]-x-[SA]-x-[LIVMT]-[KR]-x-[KR]-M-[LIVM]-[YA]-[STA](3)-x(3)-[LIVMF]-[KR]

[1] Hawkins M., Pope B., MacIver S.K., Weeds A.G. Biochemistry 32:9985-9993(1993).

30 [2] Iida K., Moriyama K., Matsumoto S., Kawasaki H., Nishida E., Yahara I. Gene 124:115-120(1993).

[3] Quirk S., MacIver S.K., Ampe C., Doberstein S.K., Kaiser D.A., van Damme J., Vandekerckhove J., Pollard T.D. Biochemistry 32:8525-8533(1993).

[4] McKim K.S., Matheson C., Marra M.A., Wakarchuk M.F., Baillie D.L. Mol. Gen. Genet. 242:346-357(1994).

[5] Moriyama K., Yonezawa N., Sakai H., Yahara I., Nishida E. J. Biol. Chem. 267:7240-7244(1992).

5

121. (Complex 24kd) Respiratory-chain NADH dehydrogenase 24 Kd subunit signature
Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or
NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner
mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as
a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this
bioenergetic enzyme complex there is one with a molecular weight of 24 Kd (in mammals),
which is a component of the iron-sulfur (IP) fragment of the enzyme. It seems to bind a 2Fe-
2S iron-sulfur cluster. The 24 Kd subunit is nuclear encoded, as a precursor form with a
transit peptide in mammals, and in *Neurospora crassa*. The 24 Kd subunit is highly similar to
[3,4]: - Subunit E of *Escherichia coli* NADH-ubiquinone oxidoreductase (gene *nuoE*). -
Subunit NQO2 of *Paracoccus denitrificans* NADH-ubiquinone oxidoreductase. A highly
conserved region, located in the central section of this subunit containing two conserved
cysteines that are probably involved in the binding of the 2Fe-2S center has been selected as a
signature pattern.

-Consensus pattern: D-x(2)-F-[ST]-x(5)-C-L-G-x-C-x(2) [GA]-P [The two C's are putative
2Fe-2S ligands]

[1] Ragan C.I. Curr. Top. Bioenerg. 15:1-36(1987).

[2] Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).

[3] Fearnley I.M., Walker J.E. Biochim. Biophys. Acta 1140:105-134(1992).

[4] Weidner U., Geier S., Ptöck A., Friedrich T., Leif H., Weiss H. J. Mol. Biol. 233:109-122(1993).

30

122. copper-bind

Copper binding proteins, plastocyanin/azurin family

Number of members: 70

Blue or 'type-1' copper proteins are small proteins which bind a single copper atom and which are characterized by an intense electronic absorption band near 600 nm [1,2]. The most well known members of this class of proteins are the plant chloroplastic plastocyanins, which exchange electrons with cytochrome c6, and the distantly related bacterial azurins, which exchange electrons with cytochrome c551. This family of proteins also includes all the proteins listed below (references are only provided for recently determined sequences).

- Amicyanin from bacteria such as *Methylobacterium extorquens* or *Thiobacillus versutus* that can grow on methylamine. Amicyanin appears to be an electron receptor for methylamine dehydrogenase.
- Auracyanins A and B from *Chloroflexus aurantiacus* [3]. These proteins can donate electrons to cytochrome c-554.
- Blue copper protein from *Alcaligenes faecalis*.
- Cupredoxin (CPC) from cucumber peelings [4].
- Cusacyanin (basic blue protein; plantacyanin, CBP) from cucumber.
- Halocyanin from *Natrobacterium pharaonis* [5], a membrane associated copper-binding protein.
- Pseudoazurin from *Pseudomonas*.
- Rusticyanin from *Thiobacillus ferrooxidans*. Rusticyanin is an electron carrier from cytochrome c-552 to the a-type oxidase [6].
- Stellacyanin from the Japanese lacquer tree.
- Umecyanin from horseradish roots.
- Allergen Ra3 from ragweed. This pollen protein is evolutionary related to the above proteins, but seems to have lost the ability to bind copper.

Although there is an appreciable amount of divergence in the sequence of all these proteins, the copper ligand sites are conserved and a pattern which includes two of the ligands (a cysteine and a histidine) has been developed.

-Consensus pattern: [GA]-x(0,2)-[YSA]-x(0,1)-[VFY]-x-C-x(1,2)-[PG]-x(0,1)-H-x(2,4)-[MQ] [C and H are copper ligands]

[1] Garret T.P.J., Clingeffer D.J., Guss J.M., Rogers S.J., Freeman H.C. J. Biol. Chem. 259:2822-2825(1984).

[2] Ryden L.G., Hunt L.T. J. Mol. Evol. 36:41-66(1993).

[3] McManus J.D., Brune D.C., Han J., Sanders-Loehr J., Meyer T.E., Cusanovich M.A., Tollin G., Blankenship R.E. J. Biol. Chem. 267:6531-6540(1992).

[4] Mann K., Schaefer W., Thoenes U., Messerschmidt A., Mehrabian Z., Nalbandyan R. FEBS Lett. 314:220-223(1992).

[5] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).

[6] Yano T., Fukumori Y., Yamanaka T. FEBS Lett. 288:159-162(1991).

123. Chaperonins cpn10 signature

Chaperonins [1,2] are proteins involved in the folding of proteins or the assembly of oligomeric protein complexes. They seem to assist other polypeptides in maintaining or assuming conformations which permit their correct assembly into oligomeric structures. They are found in abundance in prokaryotes, chloroplasts and mitochondria. Chaperonins form oligomeric complexes and are composed of two different types of subunits: a 60 Kd protein, known as cpn60 (groEL in bacteria) and a 10 Kd protein, known as cpn10 (groES in bacteria). The cpn10 protein binds to cpn60 in the presence of MgATP and suppresses the ATPase activity of the latter. Cpn10 is a protein of about 100 amino acid residues whose sequence is well conserved in bacteria, vertebrate mitochondria and plants chloroplast [3,4]. Cpn10 assembles as an heptamer that forms a dome[5]. As a signature pattern for cpn10, a region located in the N-terminal section of the protein was selected.

Consensus pattern: [LIVMFY]-x-P-[ILT]-x-[DEN]-[KR]-[LIVMFA](3)-[KREQ]-x(8,9)-[SG]-x-[LIVMFY](3)-

Note: this pattern is found twice in the plant chloroplast protein which consist of the tandem repeat of a cpn10 domain

- [1] Ellis R.J., van der Vies S.M. Annu. Rev. Biochem. 60:321-347(1991).
- [2] Zeilsta-Ryalls J., Fayet O., Georgopoulos C. Annu. Rev. Microbiol. 45:301-325(1991).
- [3] Hartman D.J., Hoogenraad N.J., Condrón R., Hoj P.B. Proc. Natl. Acad. Sci. U.S.A. 89:3394-3398(1992).
- 5 [4] Bertsch U., Soll J., Seetharam R., Viitanen P.V. Proc. Natl. Acad. Sci. U.S.A. 89:8696-8700(1992).
- [5] Hunt J.F., Weaver A.J., Landry S.J., Gierasch L., Deisenhofer J. Nature 379:37-45(1996).

10 124. Chaperonins cpn60 signature (cpn60_TCP1)

Chaperonins [1,2] are proteins involved in the folding of proteins or the assembly of oligomeric protein complexes. Their role seems to be to assist other polypeptides to maintain or assume conformations which permit their correct assembly into oligomeric structures. They are found in abundance in prokaryotes, chloroplasts and mitochondria. Chaperonins form oligomeric complexes and are composed of two different types of subunits: a 60 Kd protein, known as cpn60 (groEL in bacteria) and a 10 Kd protein, known as cpn10 (groES in bacteria). The cpn60 protein shows weak ATPase activity and is a highly conserved protein of about 550 to 580 amino acid residues which has been described by different names in different species: - Escherichia coli groEL protein, which is essential for the growth of the bacteria and the assembly of several bacteriophages. - Cyanobacterial groEL analogues. - Mycobacterium tuberculosis and leprae 65 Kd antigen, Coxiella burnetti heat shock protein B (gene htpB), Rickettsia tsutsugamushi major antigen 58, and Chlamydial 57 Kd hypersensitivity antigen (gene hypB). - Chloroplast RuBisCO subunit binding-protein alpha and beta chains, which bind ribulose biphosphate carboxylase small and large subunits and are implicated in the assembly of the enzyme oligomer. - Mammalian mitochondrial matrix protein P1 (mitonin or P60). - Yeast HSP60 protein, a mitochondrial assembly factor. As a signature pattern for these proteins, a rather well-conserved region of twelve residues, located in the last third of the cpn60 sequence was chosen.

30 Consensus pattern: A-[AS]-x-[DEQ]-E-x(4)-G-G-[GA]-

- [1] Ellis R.J., van der Vies S.M. Annu. Rev. Biochem. 60:321-347(1991).
- [2] Zeilsta-Ryalls J., Fayet O., Georgopoulos C. Annu. Rev. Microbiol. 45:301-325(1991).

Chaperonins TCP-1 signatures (cpn60_TCP1)

The TCP-1 protein [1,2] (Tailless Complex Polypeptide 1) was first identified in mice where it is especially abundant in testis but present in all cell types. It has since been found and

5 characterized in many other mammalian species, in *Drosophila* and in yeast. TCP-1 is a highly conserved protein of about 60 Kd (556 to 560 residues) which participates in a hetero-oligomeric 900 Kd double-torus shaped particle [3] with 6 to 8 other different subunits. These subunits, the chaperonin containing TCP-1 (CCT) subunit beta, gamma, delta, epsilon, zeta and eta are evolutionary related to TCP-1 itself [4,5]. The CCT is known to act as a molecular

10 chaperone for tubulin, actin and probably some other proteins. The CCT subunits are highly related to archebacterial counterparts: - TF55 and TF56 [6], a molecular chaperone from *Sulfolobus shibatae*. TF55 has ATPase activity, is known to bind unfolded polypeptides and forms an oligomeric complex of two stacked nine-membered rings. - Thermosome [7], from *Thermoplasma acidophilum*. The thermosome is composed of two subunits (alpha and beta) and also seems to be a chaperone with ATPase activity. It forms an oligomeric complex of

15 eight-membered rings. The TCP-1 family of proteins are weakly, but significantly [8], related to the cpn60/groEL chaperonin family (see <PDOC00268>). As signature patterns of this family of chaperonins, three conserved regions located in the N-terminal domain were chosen.

Consensus pattern: [RKEL]-[ST]-x-[LMFY]-G-P-x-[GSA]-x-x-K-[LIVMF](2)-

Consensus pattern: [LIVM]-[TS]-[NK]-D-[GA]-[AVNHK]-[TAV]-[LIVM](2)-x(2)-[LIVM]-x-[LIVM]-x-[SNH]-[PQH]-

Consensus pattern: Q-[DEK]-x-x-[LIVMGTA]-[GA]-D-G-T-

25 [1] Ellis J. Nature 358:191-192(1992).

[2] Nelson R.J., Craig E.A. Curr. Biol. 2:487-489(1992).

[3] Lewis V.A., Hynes G.M., Zheng D., Saibil H., Willison K.R. Nature 358:249-252(1992).

[4] Kubota H., Hynes G., Carne A., Ashworth A., Willison K.R. Curr. Biol. 4:89-99(1994)

30 [5] Kim S., Willison K.R., Horwich A.L. Trends Biochem. Sci. 20:543-548(1994).

[6] Trent J.D., Nimmesgern E., Wall J.S., Hartl F.U., Horwich A.L. Nature 354:490-493(1991).

[7] Waldmann T., Lupas A., Kellermann J., Peters J., Baumeister W. Biol. Chem. Hoppe-Seyler 376:119-126(1995).

[8] Hemmingsen S.M. Nature 357:650-650(1992).

5

125. cyclin (Cyclins)

The cyclins include an internal duplication, which is related to that found in TFIIB and the RB protein.

[1]

10

Medline: 94203808

Evidence for a protein domain superfamily shared by the cyclins, TFIIB and RB/p107.

Gibson TJ, Thompson JD, Blocker A, Kouzarides T;
Nucleic Acids Res 1994;22:946-952.

[2]

15

Medline: 96164440

The crystal structure of cyclin A

Brown NR, Noble MEM, Endicott JA, Garman EF, Wakatsuki S,
Mitchell E, Rasmussen B, Hunt T, Johnson LN;

Structure. 1995;3:1235-1247.

20

Complex of cyclin and cyclin dependant kinase.

[3]

Medline: 96313126

25

Structural basis of cyclin-dependant kinase activation by phosphorylation.

Russo AA, Jeffrey PD, Pavletich NP;
Nat Struct Biol. 1996;3:696-700.

Cyclins regulate cyclin dependant kinases (CDKs).

The most divergent prosite members have been included. Swiss:P22674

30

the Uracil-DNA glycosylase 2 is the highest noise and may be related but has not been included.

Number of members: 189

Cyclins [1,2,3] are eukaryotic proteins which play an active role in controlling nuclear cell division cycles. Cyclins, together with the p34 (cdc2) or cdk2 kinases, form the Maturation Promoting Factor (MPF). There are two main groups of cyclins:

- G2/M cyclins, essential for the control of the cell cycle at the G2/M (mitosis) transition. G2/M cyclins accumulate steadily during G2 and are abruptly destroyed as cells exit from mitosis (at the end of the M-phase).
- G1/S cyclins, essential for the control of the cell cycle at the G1/S (start) transition.

In most species, there are multiple forms of G1 and G2 cyclins. For example, in vertebrates, there are two G2 cyclins, A and B, and at least three G1 cyclins, C, D, and E.

A cyclin homolog has also been found in herpesvirus saimiri [4].

The best conserved region is in the central part of the cyclins' sequences, known as the 'cyclin-box'. From this, a 32 residue pattern has been derived.

-Consensus pattern: R-x(2)-[LIVMSA]-x(2)-[FYWS]-[LIVM]-x(8)-[LIVMFC]-x(4)-[LIVMFYA]-x(2)-[STAGC]-[LIVMFYQ]-x-[LIVMFYC]-[LIVMFY]-D-[RKH]-[LIVMFYW]

[1] Nurse P. Nature 344:503-508(1990).

[2] Norbury C., Nurse P. Curr. Biol. 1:23-24(1991).

[3] Lew D.J., Reed S.I. Trends Cell Biol. 2:77-81(1992).

[4] Nicholas J., Cameron K.R., Honess R.W. Nature 355:362-365(1992).

126. Cystatin domain

This is a very diverse family. Attempts to define separate subfamilies have failed. Typically, either the N-terminal or C-terminal end is very divergent. But splitting into two domains

would make very short families. Cathelicidins are related to this family but have not been included. Number of members: 147

Inhibitors of cysteine proteases [1,2,3], which are found in the tissues and body fluids of animals, in the larva of the worm *Onchocerca volvulus* [4], as well as in plants, can be grouped into three distinct but related families:

- Type 1 cystatins (or stefins), molecules of about 100 amino acid residues with neither disulfide bonds nor carbohydrate groups.
- Type 2 cystatins, molecules of about 115 amino acid residues which contain one or two disulfide loops near their C-terminus.
- Kininogens, which are multifunctional plasma glycoproteins.

They are the precursor of the active peptide bradykinin and play a role in blood coagulation by helping to position optimally prekallikrein and factor XI next to factor XII. They are also inhibitors of cysteine proteases. Structurally, kininogens are made of three contiguous type-2 cystatin domains, followed by an additional domain (of variable length) which contains the sequence of bradykinin. The first of the three cystatin domains seems to have lost its inhibitory activity.

In all these inhibitors, there is a conserved region of five residues which has been proposed to be important for the binding to the cysteine proteases. The consensus pattern starts one residue before this conserved region.

-Consensus pattern: [GSTEQKRV]-Q-[LIVT]-[VAF]-[SAGQ]-G-x-[LIVMNK]-x(2)-[LIVMFY]-x-[LIVMFYA]-[DENQKRHSIV]

[1] Barrett A.J. Trends Biochem. Sci. 12:193-196(1987).

[2] Rawlings N.D., Barrett A.J. J. Mol. Evol. 30:60-71(1990).

[3] Turk V., Bode W. FEBS Lett. 285:213-219(1991).

[4] Lustigman S., Brotman B., Huima T., Prince A.M. Mol. Biochem. Parasitol. 45:65-76(1991).

127. cytochrome_c (Cytochrome c)

The Pfam entry does not include all prosite members.

The cytochrome 556 and cytochrome c' families are

not included.

Number of members: 259

In proteins belonging to cytochrome c family [1], the heme group is covalently attached by thioether bonds to two conserved cysteine residues. The consensus sequence for this site is Cys-X-X-Cys-His and the histidine residue is one of the two axial ligands of the heme iron. This arrangement is shared by all proteins known to belong to cytochrome c family, which presently includes cytochromes c, c', c1 to c6, c550 to c556, cc3/Hmc, cytochrome f and reaction center cytochrome c.

-Consensus pattern: C-{CPWHF}-{CPWR}-C-H-{CFYW}

[1] Mathews F.S. Prog. Biophys. Mol. Biol. 45:1-56(1985).

128. (DAGKa) Diacylglycerol kinase accessory domain (presumed)

Diacylglycerol (DAG) is a second messenger that acts as a protein kinase C activator. This domain is assumed to be an accessory domain: its function is unknown.

[1] Sakane F, Yamada K, Kanoh H, Yokoyama C, Tanabe T, Nature 1990;344:345-348.[2] Sakane F, Imai S, Kai M, Wada I, Kanoh H, J Biol Chem 1996;271:8394-8401.[3] Schaap D, de Widt J, van der Wal J, Vandekerckhove J, van, Damme J, Gussow D, Ploegh HL, van Blitterswijk WJ, van der, Bend RL, FEBS Lett 1990;275:151-158. [4] Kanoh H, Yamada K, Sakane F, Trends Biochem Sci 1990;15:47-50.

129. (DAGKc) Diacylglycerol kinase catalytic domain (presumed)

Diacylglycerol (DAG) is a second messenger that acts as a protein kinase C activator. The catalytic domain is assumed from the finding of bacterial homologues.

[1] Sakane F, Yamada K, Kanoh H, Yokoyama C, Tanabe T, Nature 1990;344:345-348. [2] Sakane F, Imai S, Kai M, Wada I, Kanoh H, J Biol Chem 1996;271:8394-8401. [3] Schaap D, de Widt J, van der Wal J, Vandekerckhove J, van, Damme J, Gussow D, Ploegh

HL, van Blitterswijk WJ, van der, Bend RL, FEBS Lett 1990;275:151-158. [4] Kanoh H, Yamada K, Sakane F, Trends Biochem Sci 1990;15:47-50.

5 130. D-amino acid oxidases signature(DAO)

D-amino acid oxidase (EC 1.4.3.3) (DAMOX or DAO) is an FAD flavoenzyme that catalyzes the oxidation of neutral and basic D-amino acids into their corresponding keto acids. DAOs have been characterized and sequenced in fungi and vertebrates where they are known to be located in the peroxisomes. D-aspartate oxidase (EC 1.4.3.1) (DASOX) [1] is an enzyme,
10 structurally related to DAO, which catalyzes the same reaction but is active only toward dicarboxylic D-amino acids. In DAO, a conserved histidine has been shown [2] to be important for the enzyme's catalytic activity. The conserved region around this residue has been developed as a signature pattern for these enzymes.

15 Consensus pattern: [LIVM](2)-H-[NHA]-Y-G-x-[GSA](2)-x-G-x(5)-G-x-A [H is a probable active site residue]o-

[1] Negri A., Cecilian F., Tedeschi G., Simonic T., Ronchi S. J. Biol. Chem. 267:11865-11871(1992).

20 [2] Miyano M., Fukui K., Watanabe F., Takahashi S., Tada M., Kanashiro M., Miyake Y. J. Biochem. 109:171-177(1991).

131. DEAD and DEAH box families ATP-dependent helicases signatures

25 A number of eukaryotic and prokaryotic proteins have been characterized [1,2,3] on the basis of their structural similarity. They all seem to be involved in ATP-dependent, nucleic-acid unwinding. Proteins currently known to belong to this family are: - Initiation factor eIF-4A. Found in eukaryotes, this protein is a subunit of a high molecular weight complex involved in 5'cap recognition and the binding of mRNA to ribosomes. It is an ATP-dependent RNA-
30 helicase. - PRP5 and PRP28. These yeast proteins are involved in various ATP-requiring steps of the pre-mRNA splicing process. - P110, a mouse protein expressed specifically during spermatogenesis. - An3, a Xenopus putative RNA helicase, closely related to P110. - SPP81/DED1 and DBP1, two yeast proteins probably involved in pre-mRNA splicing and

related to P110. - *Caenorhabditis elegans* helicase *glh-1*. - MSS116, a yeast protein required for mitochondrial splicing. - SPB4, a yeast protein involved in the maturation of 25S ribosomal RNA. - p68, a human nuclear antigen. p68 has ATPase and DNA-helicase activities in vitro. It is involved in cell growth and division. - Rm62 (p62), a *Drosophila* putative RNA helicase related to p68. - DBP2, a yeast protein related to p68. - DHH1, a yeast protein. - DRS1, a yeast protein involved in ribosome assembly. - MAK5, a yeast protein involved in maintenance of dsRNA killer plasmid. - ROK1, a yeast protein. - *ste13*, a fission yeast protein. - Vasa, a *Drosophila* protein important for oocyte formation and specification of embryonic posterior structures. - Me31B, a *Drosophila* maternally expressed protein of unknown function. - *dbpA*, an *Escherichia coli* putative RNA helicase. - *deaD*, an *Escherichia coli* putative RNA helicase which can suppress a mutation in the *rpsB* gene for ribosomal protein S2. - *rhIB*, an *Escherichia coli* putative RNA helicase. - *rhIE*, an *Escherichia coli* putative RNA helicase. - *srmB*, an *Escherichia coli* protein that shows RNA-dependent ATPase activity. It probably interacts with 23S ribosomal RNA. - *Caenorhabditis elegans* hypothetical proteins T26G10.1, ZK512.2 and ZK686.2. - Yeast hypothetical protein YHR065c. - Yeast hypothetical protein YHR169w. - Fission yeast hypothetical protein SpAC31A2.07c. - *Bacillus subtilis* hypothetical protein *yxjN*. All these proteins share a number of conserved sequence motifs. Some of them are specific to this family while others are shared by other ATP-binding proteins or by proteins belonging to the helicases 'superfamily' [4,E1]. One of these motifs, called the 'D-E-A-D-box', represents a special version of the B motif of ATP-binding proteins. Some other proteins belong to a subfamily which have His instead of the second Asp and are thus said to be 'D-E-A-H-box' proteins [3,5,6,E1]. Proteins currently known to belong to this subfamily are: - PRP2, PRP16, PRP22 and PRP43. These yeast proteins are all involved in various ATP-requiring steps of the pre-mRNA splicing process. - Fission yeast *prh1*, which may be involved in pre-mRNA splicing. - Male-less (*mle*), a *Drosophila* protein required in males, for dosage compensation of X chromosome linked genes. - RAD3 from yeast. RAD3 is a DNA helicase involved in excision repair of DNA damaged by UV light, bulky adducts or cross-linking agents. Fission yeast *rad15* (*rhp3*) and mammalian DNA excision repair protein XPD (ERCC-2) are the homologs of RAD3. - Yeast CHL1 (or CTF1), which is important for chromosome transmission and normal cell cycle progression in G(2)/M. - Yeast TPS1. - Yeast hypothetical protein YKL078w. - *Caenorhabditis elegans* hypothetical proteins C06E1.10 and K03H1.2. - Poxviruses' early transcription factor 70 Kd subunit which acts with RNA polymerase to

initiate transcription from early gene promoters. - I8, a putative vaccinia virus helicase. - hrpA, an Escherichia coli putative RNA helicase. Signature patterns for both subfamilies were developed.

5 Consensus pattern: [LIVMF](2)-D-E-A-D-[RKEN]-x-[LIVMFYGSTN]

Consensus pattern: [GSAH]-x-[LIVMF](3)-D-E-[ALIV]-H-[NECR]

Note: proteins belonging to this family also contain a copy of the ATP/GTP- binding motif 'A' (P-loop) (see the relevant entry <[PDOC00017](#)

10 [1] Schmid S.R., Linder P. Mol. Microbiol. 6:283-292(1992).

[2] Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K., Schnier J., Slonimski P.P. Nature 337:121-122(1989).

[3] Wassarman D.A., Steitz J.A. Nature 349:463-464(1991).

[4] Hodgman T.C. Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).

15 [5] Harosh I., Deschavanne P. Nucleic Acids Res. 19:6331-6331(1991).

[6] Koonin E.V., Senkevich T.G. J. Gen. Virol. 73:989-993(1992).

132. (DHBP_synthase) 3,4-dihydroxy-2-butanone 4-phosphate synthase

20 3,4-Dihydroxy-2-butanone 4-phosphate is biosynthesized from ribulose 5-phosphate and serves as the biosynthetic precursor for the xylene ring of riboflavin. Sometimes found as a bifunctional enzyme with [GTP_cyclohydro2](#).

Richter G, Krieger C, Volk R, Kis K, Ritz H, Gotze E, Bacher A, Methods Enzymol 1997;280:374-382.

25

133. (DHDPS) Dihydrodipicolinate synthetase signatures

Dihydrodipicolinate synthetase (EC [4.2.1.52](#)) (DHDPS) [1] catalyzes, in higher plants chloroplast and in many bacteria (gene dapA), the first reaction specific to the biosynthesis of lysine and of diaminopimelate. DHDPS is responsible for the condensation of aspartate semialdehyde and pyruvate by aping-pong mechanism in which pyruvate first binds to the enzyme by forming a Schiff-base with a lysine residue. Three other proteins are structurally related to DHDPS and probably also act via a similar catalytic mechanism: - Escherichia coli

30

N-acetylneuraminate lyase (EC 4.1.3.3) (gene nanA), which catalyzes the condensation of N-acetyl-D-mannosamine and pyruvate to form N-acetylneuraminate. - Rhizobium meliloti protein mosA [3], which is involved in the biosynthesis of the rhizopine 3-o-methyl-scylo-inosamine. - Escherichia coli hypothetical protein yjhH. Two signature patterns for these enzymes were developed. The first one is centered on highly conserved region in the N-terminal part of these proteins. The second signature contains a lysine residue which has been shown, in Escherichia coli dapA [2], to be the one that forms a Schiff-base with the substrate.

Consensus pattern: [GSA]-[LIVM]-[LIVMFY]-x(2)-G-[ST]-[TG]-G-E-[GASNF]-x(6)-[EQ]

-

Consensus pattern: Y-[DNS]-[LIVMFA]-P-x(2)-[ST]-x(3)-[LIVMG]-x(13,14)-[LIVM]-x-[SGA]-[LIVMF]-K-[DEQAF]-[STAC] [K is involved in Schiff-base formation]-

[1] Kaneko T., Hashimoto T., Kumpaisal R., Yamada Y. J. Biol. Chem. 265:17451-17455(1990).

[2] Laber B., Gomis-Rueth F.-X., Romao M.J., Huber R. Biochem. J. 288:691-695(1992).

[3] Murphy P.J., Trenz S.P., Grzemeski W., de Bruijn F.J., Schell J. J. Bacteriol. 175:5193-5204 (1993).

134. (DHodehase) Dihydroorotate dehydrogenase signatures

Dihydroorotate dehydrogenase (EC 1.3.3.1) (DHodehase) catalyzes the fourth step in the de novo biosynthesis of pyrimidine, the conversion of dihydroorotate into orotate. DHodehase is a ubiquitous FAD flavoprotein. In bacteria (gene pyrD), DHodehase is located on the inner side of the cytosolic membrane. In some yeasts, such as in Saccharomyces cerevisiae (gene URA1), it is a cytosolic protein while in other eukaryotes it is found in the mitochondria [1]. The sequence of DHodehase is rather well conserved and two signature patterns were developed specific to this enzyme. The first corresponds to a region in the N-terminal section of the enzyme while the second is located in the C-terminal section and seems to be part of the FAD-binding domain.

Consensus pattern[GS]-x(4)-[GK]-[GSTA]-[LIVFSTA]-[GT]-x(3)-[NQR]-x-G-[NHY]-x(2)-P-[RT]

Consensus pattern[LIVM](2)-[GSA]-x-G-G-[IV]-x-[STGDN]-x(3)-[ACV]-x(6)-G-A

[1] Nagy M., Lacroute F., Thomas D. Proc. Natl. Acad. Sci. U.S.A. 89:8966-8970(1992).

5

135. (DMRL_synthase) 6,7-dimethyl-8-ribityllumazine synthase

136. (DNA_methylase) C-5 cytosine-specific DNA methylases signatures

10 C-5 cytosine-specific DNA methylases (EC 2.1.1.73) (C5 Mtase) are enzymes that specifically methylate the C-5 carbon of cytosines in DNA [1,2,3]. Such enzymes are found in the proteins described below. - As a component of type II restriction-modification systems in prokaryotes and some bacteriophages. Such enzymes recognize a specific DNA sequence where they methylate a cytosine. In doing so, they protect DNA from cleavage by type II restriction enzymes that recognize the same sequence. The sequences of a large number of type II C-5 Mtases are known. - In vertebrates, there are a number of C-5 Mtases that methylate CpG dinucleotides. The sequence of the mammalian enzyme is known. C-5 Mtases share a number of short conserved regions. Two of them were selected. The first is centered around a conserved Pro-Cys dipeptide in which the cysteine has been shown [4] to be involved in the catalytic mechanism; it appears to form a covalent intermediate with the C6 position of cytosine. The second region is located at the C-terminal extremity in type-II enzymes

25

Consensus pattern: [DENKS]-x-[FLIV]-x(2)-[GSTC]-x-P-C-x(2)-[FYWLIM]-S [C is the active site residue]-

Consensus pattern: [RKQGTF]-x(2)-G-N-[STAG]-[LIVMF]-x(3)-[LIVMT]-x(3)-[LIVM]-x(3)-[LIVM]-

30

[1] Posfai J., Bhagwat A.S., Roberts R.J. Gene 74:261-263(1988).

[2] Kumar S., Cheng X., Klimasauskas S., Mi S., Posfai J., Roberts R.J., Wilson G.G. Nucleic Acids Res. 22:1-10(1994).

[3] Lauster R., Trautner T.A., Noyer-Weidner M. J. Mol. Biol. 206:305-312(1989).

[4] Chen L., McMillan A.M., Chang W., Ezak-Nipkay K., Lane W.S., Verdine G.L.
 Biochemistry 30:11018-11025(1991).

5 137. (DNAphotolyase) DNA photolyases class 2 signatures

Deoxyribodipyrimidine photolyase (EC 4.1.99.3) (DNA photolyase) [1,2] is a DNA repair enzyme. It binds to UV-damaged DNA containing pyrimidine dimers and, upon absorbing a near-UV photon (300 to 500 nm), breaks the cyclobutane ring joining the two pyrimidines of the dimer. DNA photolyase is an enzyme that requires two chromophore-cofactors for its
 10 activity: a reduced FADH₂ and either 5,10-methenyltetrahydrofolate (5,10-MTFH) or an oxidized 8-hydroxy-5-deazaflavin (8-HDF) derivative (F420). The folate or deazaflavin chromophore appears to function as an antenna, while the FADH₂ chromophore is thought to be responsible for electron transfer. On the basis of sequence similarities[3] DNA
 15 photolyases can be grouped into two classes. The second class contains enzymes from *Myxococcus xanthus*, methanogenic archaeobacteria, insects, fish and marsupial mammals. It is not yet known what second cofactor is bound to class 2 enzymes. There are a number of conserved sequence regions in all known class 2 DNAphotolyases, especially in the C-terminal part. Two of these regions were selected as signature patterns.

Consensus pattern: F-x-E-E-x-[LIVM](2)-R-R-E-L-x(2)-N-F-

20 Consensus pattern: G-x-H-D-x(2)-W-x-E-R-x-[LIVM]-F-G-K-[LIVM]-R-[FY]-M-N-

[1] Sancar G.B., Sancar A. Trends Biochem. Sci. 12:259-261(1987).

[2] Jorns M.S. Biofactors 2:207-211(1990).

[3] Yasui A., Eker A.P.M., Yasuhira S., Yajima H., Kobayashi T., Takao M., Oikawa A.

25 EMBO J. 13:6143-6151(1994).

(DNAphotolyase2) DNA photolyases class 1 signatures

Deoxyribodipyrimidine photolyase (EC 4.1.99.3) (DNA photolyase) [1,2] is a DNA repair enzyme. It binds to UV-damaged DNA containing pyrimidine dimers and ,upon absorbing a
 30 near-UV photon (300 to 500 nm), breaks the cyclobutane ring joining the two pyrimidines of the dimer. DNA photolyase is an enzyme that requires two chromophore-cofactors for its activity: a reduced FADH₂ and either 5,10-methenyltetrahydrofolate (5,10-MTFH) or an oxidized 8-hydroxy-5-deazaflavin (8-HDF) derivative (F420). The folate or deazaflavin

chromophore appears to function as an antenna, while the FADH₂ chromophore is thought to be responsible for electron transfer. On the basis of sequence similarities[3] DNA photolyases can be grouped into two classes. The first class contains enzymes from Gram-negative and Gram-positive bacteria, the halophilic archaeobacteria *Halobacterium halobium*, fungi and plants. Class 1 enzymes bind either 5,10-MTHF (*E.coli*, fungi, etc.) or 8-HDF (*S.griseus*, *H.halobium*). This family also includes *Arabidopsis* cryptochromes 1 (CRY1) and 2 (CRY2), which are blue light photoreceptors that mediate blue light-induced gene expression. There are a number of conserved sequence regions in all known class 1 DNA photolyases, especially in the C-terminal part. Two of these regions were selected as signature patterns

Consensus pattern: T-G-x-P-[LIVM](2)-D-A-x-M-[RA]-x-[LIVM]-

Consensus pattern: [DN]-R-x-R-[LIVM](2)-x-[STA](2)-F-[LIVMFA]-x-K-x-L-x(2,3)- W-[KRQ]-

[1] Sancar G.B., Sancar A. Trends Biochem. Sci. 12:259-261(1987).

[2] Jorns M.S. Biofactors 2:207-211(1990).

[3] Yasui A., Eker A.P.M., Yasuhira S., Yajima H., Kobayashi T., Takao M., Oikawa A. EMBO J. 13:6143-6151(1994).

[4] Lin C., Ahmad M., Cashmore A.R. Plant J. 10:893-902(1996).

138. (DNA_pol_A)

DNA polymerase family A signature

Replicative DNA polymerases (EC 2.7.7.7) are the key enzymes catalyzing the accurate replication of DNA. They require either a small RNA molecule or a protein as a primer for the de novo synthesis of a DNA chain. On the basis of sequence similarities a number of DNA polymerases have been grouped together [1,2,3] under the designation of DNA polymerase family A. The polymerases that belong to this family are listed below.

- *Escherichia coli* and various other bacterial polymerase I (gene polA).
- *Thermus aquaticus* Taq polymerase.
- Bacteriophage sp01 polymerase.

- Bacteriophage sp02 polymerase.
- Bacteriophage T5 polymerase.
- Bacteriophage T7 polymerase.
- Mycobacteriophage L5 polymerase.
- 5 - Yeast mitochondrial polymerase gamma (gene MIP1).

Five regions of similarity are found in all the above polymerases. One of these conserved regions, known as 'motif B' [1], is located in a domain which, in *Escherichia coli* polA, has been shown to bind deoxynucleotide triphosphate substrates; it contains a conserved tyrosine which has been shown, by photo- affinity labelling, to be in the active site; a conserved lysine, also part of this motif, can be chemically labelled, using pyridoxal phosphate. This conserved region was used as a signature for this family of DNA polymerases.

Consensus pattern R-x(2)-[GSAV]-K-x(3)-[LIVMFY]-[AGQ]-x(2)-Y-x(2)-[GS]-x(3)-[LIVMA] Sequences known to belong to this class detected by the pattern ALL.

- [1] Delarue M., Poch O., Todro N., Moras D., Argos P. *Protein Eng.* 3:461-467(1990).
- [2] Ito J., Braithwaite D.K. *Nucleic Acids Res.* 19:4045-4057(1991).
- [3] Braithwaite D.K., Ito J. *Nucleic Acids Res.* 21:787-802(1993).

139. DNA_pol_viral_C

DNA polymerase (viral) C-terminal domain

Number of members: 128

140. (DNA_topoisII)

DNA topoisomerase II signature

DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type II topoisomerases are ATP-dependent and act by passing a DNA segment through a transient double-strand break.

Topoisomerase II is found in phages, archaeobacteria, prokaryotes, eukaryotes, and in African Swine Fever virus (ASF). In bacteriophage T4 topoisomerase II consists of three subunits

(the product of genes 39, 52 and 60). In prokaryotes and in archaeobacteria the enzyme, known as DNA gyrase, consists of two subunits (genes *gyrA* and *gyrB* [E2]). In some bacteria, a second type II topoisomerase has been identified; it is known as topoisomerase IV and is required for chromosome segregation, it also consists of two subunits (genes *parC* and *parE*). In eukaryotes, type II topoisomerase is a homodimer.

There are many regions of sequence homology between the different subtypes of topoisomerase II. The relation between the different subunits is shown in the following representation:

```

<-----About-1400-residues----->

[-----Protein 39-*-----][----Protein 52----]      Phage T4
[-----gyrB-----*-----][-----gyrA-----]  Prokaryote II
                                     Archaeobacteria
[-----parE-----*-----][-----parD-----]  Prokaryote IV
[-----*-----] Eukaryote and
                                     ASF

```

'*': Position of the pattern.

As a signature pattern for this family of proteins, a region that contains a highly conserved pentapeptide was selected. The pattern is located in *gyrB*, in *parE*, and in protein 39 of phage T4 topoisomerase.

Consensus pattern[LIVMA]-x-E-G-[DN]-S-A-x-[STAG] Sequences known to belong to this class detected by the pattern ALL.

[1] Sternglanz R. Curr. Opin. Cell Biol. 1:533-535(1990).

[2] Bjornsti M.-A. Curr. Opin. Struct. Biol. 1:99-103(1991).

[3] Sharma A., Mondragon A. Curr. Opin. Struct. Biol. 5:39-47(1995).

[4] Roca J. Trends Biochem. Sci. 20:156-160(1995).

141. (DSPc) Tyrosine specific protein phosphatases signature and profiles

Tyrosine specific protein phosphatases (EC 3.1.3.48) (PTPase) [1 to 5] are enzymes that catalyze the removal of a phosphate group attached to a tyrosine residue. These enzymes are very important in the control of cell growth, proliferation, differentiation and transformation.

Multiple forms of PTPase have been characterized and can be classified into two categories: soluble PTPases and transmembrane receptor proteins that contain PTPase domain(s). The currently known PTPases are listed below: Soluble PTPases. - PTPN1 (PTP-1B). - PTPN2 (T-cell PTPase; TC-PTP). - PTPN3 (H1) and PTPN4 (MEG), enzymes that contain an N-terminal band 4.1- like domain (see <PDOC00566>) and could act at junctions between the membrane and cytoskeleton. - PTPN5 (STEP). - PTPN6 (PTP-1C; HCP; SHP) and PTPN11 (PTP-2C; SH-PTP3; Syp), enzymes which contain two copies of the SH2 domain at its N-terminal extremity. The Drosophila protein corkscrew (gene csw) also belongs to this subgroup. - PTPN7 (LC-PTP; Hematopoietic protein-tyrosine phosphatase; HePTP). - PTPN8 (70Z-PEP). - PTPN9 (MEG2). - PTPN12 (PTP-G1; PTP-P19). - Yeast PTP1. - Yeast PTP2 which may be involved in the ubiquitin-mediated protein degradation pathway. - Fission yeast pyp1 and pyp2 which play a role in inhibiting the onset of mitosis. - Fission yeast pyp3 which contributes to the dephosphorylation of cdc2. - Yeast CDC14 which may be involved in chromosome segregation. - Yersinia virulence plasmid PTPases (gene yopH). - Autographa californica nuclear polyhedrosis virus 19 Kd PTPase. Dual specificity PTPases. - DUSP1 (PTPN10; MAP kinase phosphatase-1; MKP-1); which dephosphorylates MAP kinase on both Thr-183 and Tyr-185. - DUSP2 (PAC-1), a nuclear enzyme that dephosphorylates MAP kinases ERK1 and ERK2 on both Thr and Tyr residues. - DUSP3 (VHR). - DUSP4 (HVH2). - DUSP5 (HVH3). - DUSP6 (Pyst1; MKP-3). - DUSP7 (Pyst2; MKP-X). - Yeast MSG5, a PTPase that dephosphorylates MAP kinase FUS3. - Yeast YVH1. - Vaccinia virus H1 PTPase; a dual specificity phosphatase. Receptor PTPases. Structurally, all known receptor PTPases, are made up of a variable length extracellular domain, followed by a transmembrane region and a C-terminal catalytic cytoplasmic domain. Some of the receptor PTPases contain fibronectin type III (FN-III) repeats, immunoglobulin-like domains, MAM domains or carbonic anhydrase-like domains in their extracellular region. The cytoplasmic region generally contains two copies of the PTPase domain. The first seems to have enzymatic activity, while the second is inactive but seems to affect substrate specificity of the first. In these domains, the catalytic cysteine is generally conserved but some other, presumably important, residues are not. In the following table, the domain structure of known

receptor PTPases is shown: Extracellular Intracellular ----- Ig FN-3
 CAH MAM PTPaseLeukocyte common antigen (LCA) (CD45) 0 2 0 0 2Leukocyte antigen
 related (LAR) 3 8 0 0 2 Drosophila DLAR 3 9 0 0 2Drosophila DPTP 2 2 0 0 2PTP-alpha
 (LRP) 0 0 0 0 2PTP-beta 0 16 0 0 1PTP-gamma 0 1 1 0 2PTP-delta 0 >7 0 0 2 PTP-epsilon 0
 5 0 0 0 2PTP-kappa 1 4 0 1 2PTP-mu 1 4 0 1 2PTP-zeta 0 1 1 0 2PTPase domains consist of
 about 300 amino acids. There are two conserved cysteines, the second one has been shown to
 be absolutely required for activity. Furthermore, a number of conserved residues in its
 immediate vicinity have also been shown to be important. A signature pattern for PTPase
 domains was derived centered on the active site cysteine. There are three profiles for
 10 PTPases, the first one spans the complete domain and is not specific to any subtype. The
 second profile is specific to dual-specificity PTPases and the third one to the PTP subfamily

Consensus pattern: [LIVMF]-H-C-x(2)-G-x(3)-[STC]-[STAGP]-x-[LIVMFY] [C is the
 active site residue]-

- [1] Fischer E.H., Charbonneau H., Tonks N.K. Science 253:401-406(1991).
- [2] Charbonneau H., Tonks N.K. Annu. Rev. Cell Biol. 8:463-493(1992).
- [3] Trowbridge I.S. J. Biol. Chem. 266:23517-23520(1991).
- [4] Tonks N.K., Charbonneau H. Trends Biochem. Sci. 14:497-500(1989).
- [5] Hunter T. Cell 58:1013-1016(1989).

142. (DUF10) Uncharacterized protein family UPF0076 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -

- Goat antigen UK114, a human homolog and the rat corresponding protein which is known as
 perchloric acid soluble protein (PSP1). PSP1 [2] may inhibit an initiation stage of cell-free
 protein synthesis. - Mouse heat-responsive protein HRSP12. - Yeast chromosome V
 hypothetical protein YER057c. - Yeast chromosome IX hypothetical protein YIL051c. -
 Caenorhabditis elegans hypothetical protein C23G10.2. - Escherichia coli hypothetical
 protein ycdK. - Escherichia coli hypothetical protein yhaR. - Escherichia coli hypothetical
 protein yjgF and HI0719, the corresponding Haemophilus influenzae protein. - Escherichia
 coli hypothetical protein yoaB. - Bacillus subtilis hypothetical protein yabJ. - Haemophilus
 influenzae hypothetical protein HI1627. - Helicobacter pylori hypothetical protein HP0944. -

Lactococcus lactis aldR. - Myxococcus xanthus dfrA. - Synechocystis strain PCC 6803
hypothetical protein slr0709. - Rhizobium strain NGR234 symbiotic plasmid hypothetical
protein y4sK. - Pyrococcus horikoshii hypothetical protein PH0854. These are small proteins
of around 15 Kd whose sequence is highly conserved. As a signature pattern, a well conserved
5 region located in the C-terminal part of these proteins was selected.

Consensus pattern: [PA]-[ASTPV]-R-[SACVF]-x-[LIVMFY]-x(2)-[GSAKR]-x-[LMVA]-
x(5,8)-[LIVM]-E-[MI]-

10 [1] Bairoch A. Unpublished observations (1995).
[2] Oka T., Tsuji H., Noda C., Sakai K., Hong Y.-M., Suzuki I., Munoz S., Natori Y. J. Biol.
Chem. 270:30060-30067(1995).

15 143. (DUF3) Domain of Unknown Function 3
Domain apparently occurring exclusively in eubacteria. Unknown
function.

20 144. (DUF6) Integral membrane protein
This family includes many hypothetical membrane proteins of unknown function.
Many of the proteins contain two copies of the aligned region.

25 145. (DUF7) Integral membrane protein
This family includes many hypothetical membrane proteins of unknown function.
Swiss:P14502 has been implicated in resistance to ethidium bromide.

30 146. (DapB) Dihydrodipicolinate reductase signature
Dihydrodipicolinate reductase (EC 1.3.1.26) catalyzes the second step in the biosynthesis of
diaminopimelic acid and lysine, the NAD or NADP-dependent reduction of 2,3-
dihydrodipicolinate into 2,3,4,5-tetrahydrodipicolinate. This enzyme is present in bacteria

(gene *dapB*) and higher plants. As a signature pattern the best conserved region in this enzyme was selected. It is located in the central section and is part of the substrate-binding region [1].

5 Consensus pattern: E-[IV]-x-E-x-H-x(3)-K-x-D-x-P-S-G-T-A-

[1] Scapin G., Blanchard J.S., Sacchettini J.C. Biochemistry 34:3502-3512(1995).

10 147. DedA family

This family combines the DedA related proteins and YIAN/YGIK family. Members of this family are not functionally characterised. These proteins contain multiple predicted transmembrane regions.

15 148. DegT/DnrJ/EryC1/StrS family

The members of this family exhibit some characteristics of the sensor protein of two-component signal transduction systems, however none of the members show any sequence similarity to these protein kinases. The members of this family do have the typical helix-turn-helix motif of DNA binding proteins.

[1] Stutzman-Engwall KJ, Otten SL, Hutchinson CR, J Bacteriol 1992;174:144-154.

20 149. (Desaturase) Fatty acid desaturases signatures

25 Fatty acid desaturases (EC 1.14.99.-) are enzymes that catalyze the insertion of a double bond at the delta position of fatty acids. There seems to be two distinct families of fatty acid desaturases which do not seem to be evolutionary related. Family 1 is composed of: - Stearoyl-CoA desaturase (SCD) (EC 1.14.99.5) [1]. SCD is a key regulatory enzyme of unsaturated fatty acid biosynthesis. SCD introduces a cis double bond at the delta(9) position of fatty acyl-CoA's such as palmitoleoyl- and oleoyl-CoA. SCD is a membrane-bound enzyme that is thought to function as a part of a multienzyme complex in the endoplasmic reticulum of vertebrates and fungi. As a signature pattern for this family a conserved region in the C-terminal part of these enzymes was selected, this region is rich in histidine residues

30

and in aromatic residues. Family 2 is composed of: - Plants stearyl-acyl-carrier-protein desaturase (EC 1.14.99.6) [2], these enzymes catalyze the introduction of a double bond at the delta(9) position of stearyl-ACP to produce oleoyl-ACP. This enzyme is responsible for the conversion of saturated fatty acids to unsaturated fatty acids in the synthesis of vegetable oils. - Cyanobacteria desA [3] an enzyme that can introduce a second cis double bond at the delta(12) position of fatty acid bound to membranes glycerolipids. DesA is involved in chilling tolerance; the phase transition temperature of lipids of cellular membranes being dependent on the degree of unsaturation of fatty acids of the membrane lipids. As a signature pattern for this family a conserved region in the C-terminal part of these enzymes was selected.

Consensus pattern: G-E-x-[FY]-H-N-[FY]-H-H-x-F-P-x-D-Y-

Consensus pattern: [ST]-[SA]-x(3)-[QR]-[LI]-x(5,6)-D-Y-x(2)-[LIVMFYW]-[LIVM]- [DE]-

[1] Kaestner K.H., Ntambi J.M., Kelly T.J. Jr., Lane M.D. J. Biol. Chem. 264:14755-14761(1989).

[2] Shanklin J., Somerville C.R. Proc. Natl. Acad. Sci. U.S.A. 88:2510-2514(1991).

[3] Wada H., Gombos Z., Murata N. Nature 347:200-203(1990).

150. Dihydroorotase signatures

Dihydroorotase (EC 3.5.2.3) (DHOase) catalyzes the third step in the de novo biosynthesis of pyrimidine, the conversion of ureidosuccinic acid (N-carbamoyl-L-aspartate) into dihydroorotate. Dihydroorotase binds a zinc ion which is required for its catalytic activity [1].

In bacteria, DHOase is a dimer of identical chains of about 400 amino-acid residues (gene pyrC). In higher eukaryotes, DHOase is part of a large multi-functional protein known as 'rudimentary' in Drosophila and CAD in mammals and which catalyzes the first three steps of pyrimidine biosynthesis [2]. The DHOase domain is located in the central part of this polyprotein. In yeasts, DHOase is encoded by a monofunctional protein (gene URA4).

However, a defective DHOase domain [3] is found in a multifunctional protein (gene URA2) that catalyzes the first two steps of pyrimidine biosynthesis. The comparison of DHOase sequences from various sources shows [4] that there are two highly conserved regions. The first located in the N-terminal extremity contains two histidine residues

suggested [3] to be involved in binding the zinc ion. The second is found in the C-terminal part. Signature patterns for both regions have been developed. Allantoinase (EC 3.5.2.5) is the enzyme that hydrolyzes allantoin into allantoate. In yeast (gene DAL1) [5], it is the first enzyme in the allantoin degradation pathway; in amphibians [6] and fish it catalyzes the second step in the degradation of uric acid. The sequence of allantoinase is evolutionary related to that of DHOases.

Consensus pattern: D-[LIVMFYWSAP]-H-[LIVA]-H-[LIVF]-[RN]-x-[PGANF] [The two H's are probable zinc ligands]-

Consensus pattern: [GA]-[ST]-D-x-A-P-H-x(4)-K-

[1] Brown D.C., Collins K.D. J. Biol. Chem. 266:1597-1604(1991).

[2] Davidson J.N., Chen K.C., Jamison R.S., Musmanno L.A., Kern C.B. BioEssays 15:157-164(1993).

[3] Souciet J.-L., Nagy M., Le Gouar M., Lacroute F., Potier S. Gene 79:59-70(1989).

[4] Guyonvarch A., Nguyen-Juilleret M., Hubert J.-C., Lacroute F. Mol. Gen. Genet. 212:134-141(1988).

[5] Buckholz R.G., Cooper T.G. Yeast 7:913-923(1991).

[6] Hayashi S., Jain S., Chu R., Alvares K., Xu B., Erfurth F., Usuda N., Rao M.S., Reddy S.K., Noguchi T., Reddy J.K., Yeldandi A.Y. J. Biol. Chem. 269:12269-12276(1994).

151. dnaJ domains signatures and profile

The prokaryotic heat shock protein dnaJ interacts with the chaperone hsp70-like dnaK protein [1]. Structurally, the dnaJ protein consists of an N-terminal conserved domain (called 'J' domain) of about 70 amino acids, a glycine-rich region ('G' domain') of about 30 residues, a central domain containing four repeats of a CXXCXGXG motif ('CRR' domain) and a C-terminal region of 120 to 170 residues. Such a structure is shown in the following schematic representation:

```

+-----+-----+-----+-----+-----+-----+-----+-----+ | N-terminal | |
Gly-R | | CXXCXGXG | C-terminal | +-----+-----+-----+-----+
-----+
```

It has been shown [2] that the 'J' domain as well as the 'CRR' domain are also found in other prokaryotic and eukaryotic proteins which are listed below.

a) Proteins containing both a 'J' and a 'CRR' domain:

- Yeast protein MAS5/YDJ1 which seems to be involved in mitochondrial protein import.
- Yeast protein MDJ1, involved in mitochondrial biogenesis and protein folding.
- Yeast protein SCJ1, involved in protein sorting.
- Yeast protein XDJ1.
- Plants dnaJ homologs (from leek and cucumber).
- Human HDJ2, a dnaJ homolog of unknown function.
- Yeast hypothetical protein YNL077w.

b) Proteins containing a 'J' domain without a 'CRR' domain:

- Rhizobium fredii nolC, a protein involved in cultivar-specific nodulation of soybean.
- Escherichia coli cbpA [3], a protein that binds curved DNA.
- Yeast protein SEC63/NPL1, important for protein assembly into the endoplasmic reticulum and the nucleus.
- Yeast protein SIS1, required for nuclear migration during mitosis.
- Yeast protein CAJ1.
- Yeast hypothetical protein YFR041c.
- Yeast hypothetical protein YIR004w.
- Yeast hypothetical protein YJL162c.
- Plasmodium falciparum ring-infected erythrocyte surface antigen (RESA). RESA, whose function is not known, is associated with the membrane skeleton of newly invaded erythrocytes.
- Human HDJ1.
- Human HSJ1, a neuronal protein.
- Drosophila cysteine-string protein (csp).

A signature pattern for the 'J' domain was developed, based on conserved positions in the C-terminal half of this domain. A pattern for the 'CRR' domain, based on the first two copies of that motif was also developed. A profile for the 'J' domain was also developed.

Consensus pattern: [FY]-x(2)-[LIVMA]-x(3)-[FYWHNT]-[DENQSA]-x-L-x-[DN]-x(3)-
[KR]-x(2)-[FYI]-

Consensus pattern: C-[DEGSTHKK]-x-C-x-G-x-[GK]-[AGSDM]-x(2)-[GSNKR]-x(4,6)-C-
x(2,3)-C-x-G-x-G-

5

[1] Cyr D.M., Langer T., Douglas M.G. Trends Biochem. Sci. 19:176-181(1994).

[2] Bork P., Sander C., Valencia A., Bukau B. Trends Biochem. Sci. 17:129-129(1992).

[3] Ueguchi C., Kaneda M., Yamada H., Mizuno T. Proc. Natl. Acad. Sci. U.S.A. 91:1054-
1058(1994).

10

152.

153. Dwarfins

This family known as the dwarfins also includes the drosophila protein MAD. The N-
terminus of MAD can bind to DNA [2].

[1] Yingling JM, Das P, Savage C, Zhang M, Padgett RW, Wang XF, Proc Natl Acad
Sci U S A 1996;93:8940-8944. [2] Kim J, Johnson K, Chen HJ, Carroll S, Laughon A,
Nature 1997;388:304-308.

154. Dynein light chain type 1 signature

Dynein is a multisubunit microtubule-dependent motor enzyme that acts as the force
generating protein of eukaryotic cilia and flagella. The cytoplasmic isoform of dynein acts as
a motor for the intracellular retrograde motility of vesicles and organelles along microtubules.
Dynein is composed of a number of ATP-binding large subunits, intermediate size subunits
and small subunits. Among the small subunits, there is a family [1,2] of highly conserved
proteins which consist of: - Chlamydomonas reinhardtii flagellar outer arm dynein 8 Kd and
11 Kd light chains. - Higher eukaryotes cytoplasmic dynein light chain 1. - Yeast cytoplasmic
dynein light chain 1 (gene DYN2 or SLC1). - Caenorhabditis elegans hypothetical dynein
light chains M18.2 and T26A5.9. These proteins have from 89 to 120 amino acids. As a
signature pattern, A highly conserved region was selected.

25

30

Consensus pattern: H-x-I-x-G-[KR]-x-F-[GA]-S-x-V-[ST]-[HY]-E -

[1] King S.M., Patel-King R.S. *J. Biol. Chem.* 270:11445-11452(1995).

[2] Dick T., Ray K., Salz H.K., Chia W. *Mol. Cell. Biol.* 16:1966-1977(1996).

5

155. dUTPase

dUTPase hydrolyzes dUTP to dUMP and pyrophosphate.

[1] Cedergren-Zeppezauer ES, Larsson G, Nyman PO, Dauter Z, Wilson KS, *Nature* 1992;355:740-743. [2] Mol CD, Harris JM, McIntosh EM, Tainer JA, *Structure* 1996;4:1077-1092.

15
20
25
30

156. (dCMP cyt deam) Cytidine and deoxycytidylate deaminases zinc-binding region signature

Cytidine deaminase (EC 3.5.4.5) (cytidine aminohydrolase) catalyzes the hydrolysis of cytidine into uridine and ammonia while deoxycytidylatedeaminase (EC 3.5.4.12) (dCMP deaminase) hydrolyzes dCMP into dUMP. Both enzymes are known to bind zinc and to require it for their catalytic activity[1,2]. These two enzymes do not share any sequence similarity with the exception of a region that contains three conserved histidine and cysteine residues which are thought to be involved in the binding of the catalytic zincion. Such a region is also found in other proteins [3,4]: - Yeast cytosine deaminase (EC 3.5.4.1) (gene FCY1) which transforms cytosine into uracil. - Mammalian apolipoprotein B mRNA editing protein, responsible for the postranscriptional editing of a CAA codon into a UAA (stop) codon in the APOB mRNA. - Riboflavin biosynthesis protein ribG, which converts 2,5-diamino-6- (ribosylamino)-4(3H)-pyrimidinone 5'-phosphate into 5-amino-6- (ribosylamino)-2,4(1H,3H)-pyrimidinedione 5'-phosphate. - *Bacillus cereus* blasticidin-S deaminase (EC 3.5.4.23), which catalyzes the deamination of the cytosine moiety of the antibiotics blasticidin S, cytomycin and acetylblasticidin S. - *Bacillus subtilis* protein comEB. This protein is required for the binding and uptake of transforming DNA. - *Bacillus subtilis* hypothetical protein yaaJ. - *Escherichia coli* hypothetical protein yfhC. - Yeast hypothetical protein YJL035c. A signature pattern for this zinc-binding region was derived.

Consensus pattern: [CH]-[AGV]-E-x(2)-[LIVMFGAT]-[LIVM]-x(17,33)-P-C-x(2,8)-C-x(3)-[LIVM] [The C's and H are zinc ligands]

- [1] Yang C., Carlow D., Wolfenden R., Short S.A. Biochemistry 31:4168-4174(1992).
 [2] Moore J.T., Silversmith R.E., Maley G.F., Maley F. J. Biol. Chem. 268:2288-2291(1993).
 [3] Reizer J., Buskirk S., Bairoch A., Reizer A., Saier M.H. Jr. Protein Sci. 3:853-856(1994).
 [4] Bhattacharya S., Navaratnam N., Morrison J.R., Scott J., Taylow W.R. Trends Biochem. Sci. 19:105-106(1994).

157. Dehydrins signatures

A number of proteins are produced by plants that experience water-stress. Water-stress takes place when the water available to a plant falls below a critical level. The plant hormone abscisic acid (ABA) appears to modulate the response of plant to water-stress. Proteins that are expressed during water-stress are called dehydrins [1,2] or LEA group 2 proteins [3]. The proteins that belong to this family are listed below.

- Arabidopsis thaliana XERO 1, XERO 2 (LTI30), RAB18, ERD10 (LTI45) ERD14 and COR47.
- Barley dehydrins B8, B9, B17, and B18.
- Cotton LEA protein D-11.
- Craterostigma plantagineum dessication-related proteins A and B.
- Maize dehydrin M3 (RAB-17).
- Pea dehydrins DHN1, DHN2, and DHN3.
- Radish LEA protein.
- Rice proteins RAB 16B, 16C, 16D, RAB21, and RAB25.
- Tomato TAS14.
- Wheat dehydrin RAB 15 and cold-shock protein cor410, cs66 and cs120.

Dehydrins share a number of structural features. One of the most notable features is the presence, in their central region, of a continuous run of five to nine serines followed by a cluster of charged residues. Such a region has been found in all known dehydrins so far with the exception of pea dehydrins. A second conserved feature is the presence of two copies of lysine-rich octapeptide; the first copy is located just after the cluster of charged residues that follows the poly-serine region and the second copy is found at the C-terminal extremity. Signature patterns for both regions were derived.

Consensus pattern: S(5)-[DE]-x-[DE]-G-x(1,2)-G-x(0,1)-[KR](4

Consensus pattern: [KR]-[LIM]-K-[DE]-K-[LIM]-P-G-

[1] Close T.J., Kortt A.A., Chandler P.M. Plant Mol. Biol. 13:95-108(1989).

[2] Robertson M., Chandler P.M. Plant Mol. Biol. 19:1031-1044(1992).

5 [3] Dure L. III, Crouch M., Harada J., Ho T.-H. D., Mundy J., Quatrano R., Thomas T., Sung Z.R. Plant Mol. Biol. 12:475-486(1989).

158. (deoR) Bacterial regulatory proteins, deoR family signature

10 The many bacterial transcription regulation proteins which bind DNA through a helix-turn-helix' motif can be classified into subfamilies on the basis of sequence similarities. One of these subfamilies groups the following proteins[1,2]: - accR, the Agrobacterium tumefaciens plasmid pTiC58 repressor of opine catabolism and conjugal transfer. - agaR, the Escherichia coli aga operon putative repressor. - deoR, the Escherichia coli deoxyribose operon repressor. - fucR, the Escherichia coli L-fucose operon activator. - gatR, the Escherichia coli galactitol operon repressor. - glpR, the Escherichia coli glycerol-3-phosphate regulon repressor. - gutR (or srlR), the Escherichia coli glucitol operon repressor. - iolR, from Bacillus subtilis. - lacR, the streptococci lactose phosphotransferase system repressor. - spoIIID, the Bacillus subtilis transcription regulator of the sigK gene. - yfjR, an Escherichia coli hypothetical protein. - ygbI, an Escherichia coli hypothetical protein. - yihW, an Escherichia coli hypothetical protein. - yjfQ, an Escherichia coli hypothetical protein. - yjhJ, an Escherichia coli hypothetical protein. The 'helix-turn-helix' DNA-binding motif of these proteins is located in the N-terminal part of the sequence. The pattern used to detect these proteins starts fourteen residues before the HTH motif and ends one residue after it.

25 Consensus pattern: R-x(3)-[LIVM]-x(3)-[LIVM]-x(16,17)-[STA]-x(2)-T-[LIVMA]- [RH]-[KRNA]-D-[LIVMF]-

[1] von Bodman S., Hayman G.T., Farrand S.K. Proc. Natl. Acad. Sci. U.S.A. 89:643-647(1992).

[2] Bairoch A. Unpublished observations (1993).

159. dsrm

Double-stranded RNA binding motif

[1] Burd CG, Dreyfuss G; Medline: 94310455, Conserved structures and diversity of
5 functions of RNA-binding proteins. Science 1994;265:615-621.

Sequences gathered for seed by HMM_iterative_training Putative motif shared by proteins
that bind to dsRNA. At least some DSRM proteins seem to bind to specific RNA targets.
Exemplified by Staufen, which is involved in localization of at least five different mRNAs in
10 the early Drosophila embryo. Also by interferon-induced protein kinase in humans, which is
part of the cellular response to dsRNA.

Number of members: 116

160. Dynamin family signature

Dynamin [1,2] is a microtubule-associated force-producing protein of 100 Kd which is
involved in the production of microtubule bundles and which is able to bind and hydrolyze
GTP. Dynamin is structurally related to the following proteins: - Drosophila shibire protein
(gene shi) [3]. Shibire is, very probably, the Drosophila cognate of mammalian dynamin. It
seems to provide the motor for vesicular transport during endocytosis. - Yeast vacuolar
sorting protein VPS1 (or SPO15) [4], a protein which could also be involved in microtubule-
associated motility. - Yeast protein MGM1 [5], which is required for mitochondrial genome
maintenance. - Yeast protein DNM1, which is involved in endocytosis. - Interferon induced
25 Mx proteins [6,7]. Interferon alpha or beta induce the synthesis of a family of closely related
proteins. Most of these proteins are known to confer resistance to influenza viruses and/or
rhabdoviruses on transfected mammalian cell in culture. The three motifs found in all GTP-
binding proteins are located in the N-terminal part of these proteins. The signature pattern
that was developed for these proteins is based on a highly conserved region downstream of
30 the ATP/GTP-binding motif 'A' (P-loop) (see <[PDOC00017](#)>).-

Consensus pattern: L-P-[RK]-G-[STN]-[GN]-[LIVM]-V-T-R-

- [1] Vallee R.B., Shpetner H.S. Annu. Rev. Biochem. 59:909-932(1990).
- [2] Obar R.A., Collins C.A., Hammarback J.A., Shpetner H.S., Vallee R.B. Nature 347:256-261(1990).
- [3] van der Blik A., Meyerowitz E.M. Nature 351:411-414(1991).
- 5 [4] Rothman J.H., Raymond C.K., Gilbert T., O'Hara P.J., Stevens T.H. Cell 61:1063-1074(1990).
- [5] Jones B.A., Fangman W.L. Genes Dev. 6:380-389(1992).
- [6] Arnheiter H., Meier E. New Biol. 2:851-857(1990).
- [7] Staeheli P., Pitossi F., Pavlovic J. Trends Cell Biol. 3:268-272(1993).

10

161. (dynamin_2) Dynamin central region

This region lies between the GTPase domain, see dynamin, and the pleckstrin homology (PH) domain.

15

162. E1-E2 ATPases phosphorylation site

E1-E2 ATPases (also known as P-type) are cation transport ATPases which form an aspartyl phosphate intermediate in the course of ATP hydrolysis. ATPases which belong to this family are listed below [1,2,3]. - Fungal and plant plasma membrane (H⁺) ATPases [reviewed in 4]. - Vertebrate (Na⁺, K⁺) ATPases (sodium pump) [reviewed in 5,6]. - Gastric (K⁺, H⁺) ATPases (proton pump). - Calcium (Ca⁺⁺) ATPases (calcium pump) from the sarcoplasmic reticulum (SR), the endoplasmic reticulum (ER) and the plasma membrane. - Copper (Cu⁺⁺) ATPases (copper pump) which are involved in two human genetic disorders: Menkes syndrome and Wilson disease [7]. - Bacterial potassium (K⁺) ATPases. - Bacterial cadmium efflux (Cd⁺⁺) ATPases [reviewed in 8]. - Bacterial magnesium (Mg⁺⁺) ATPases. - A probable cation ATPase from Leishmania. - fixI, a probable cation ATPase from Rhizobium meliloti, involved in nitrogen fixation. The region around the phosphorylated aspartate residue is perfectly conserved in all these ATPases and can be used as a signature pattern.

20

Consensus pattern: D-K-T-G-T-[LI]-[TI] [D is phosphorylated]

30

- [1] Green N.M., McLennan D.H. Biochem. Soc. Trans. 17:819-822(1989).

- [2] Green N.M. Biochem. Soc. Trans. 17:970-972(1989).
- [3] Fagan M.J., Saier M.H. Jr. J. Mol. Evol. 38:57-99(1994).
- [4] Serrano R. Biochim. Biophys. Acta 947:1-28(1988).
- [5] Fambrough D.M. Trends Neurosci. 11:325-328(1988).
- 5 [6] Sweadner K.J. Biochim. Biophys. Acta 988:185-220(1989).
- [7] Bull P.C., Cox D.W. Trends Genet. 10:246-251(1994).
- [8] Silver S., Nucifora G., Chu L., Misra T.K. Trends Biochem. Sci. 14:76-80(1989).

10 163. E1_N

E1 Protein, N terminal domain

Number of members: 90

15 164. (E1_dehydrog) Dehydrogenase E1 component

This family uses thiamine pyrophosphate as a cofactor. This family includes pyruvate dehydrogenase, 2-oxoglutarate dehydrogenase and 2-oxoisovalerate dehydrogenase.

20 165. (ECH) Enoyl-CoA hydratase/isomerase signature

Enoyl-CoA hydratase (EC 4.2.1.17) (ECH) [1] and 3-2trans-enoyl-CoA isomerase(EC 5.3.3.8) (ECI) [2] are two enzymes involved in fatty acid metabolism. ECH catalyzes the hydration of 2-trans-enoyl-CoA into 3-hydroxyacyl-CoA and ECI shifts the 3- double bond of the intermediates of unsaturated fatty acid oxidation to the 2-trans position. Most eukaryotic cells have two fatty-acid beta-oxidation systems, one located in mitochondria and the other in peroxisomes. In mitochondria, ECH and ECI are separate yet structurally related monofunctional enzymes. Peroxisomes contain a trifunctional enzyme [3] consisting of an N-terminal domain that bears both ECH and ECI activity, and a C-terminal domain responsible for 3-hydroxyacyl-CoA dehydrogenase (HCDH) activity. In Escherichia coli (gene fadB) and Pseudomonas fragi (gene faoA), ECH and ECI are also part of a multifunctional enzyme which contains both a HCDH and a3-hydroxybutyryl-CoA epimerase domain [4].A number of other proteins have been found to be evolutionary related to the ECH/ECI enzymes or domains: - 3-hydroxybutyryl-coa dehydratase (EC 4.2.1.55) (crotonase), a bacterial enzyme

involved in the butyrate/butanol-producing pathway. - Naphthoate synthase (EC 4.1.3.36) (DHNA synthetase) (gene *menB*) [5], a bacterial enzyme involved in the biosynthesis of menaquinone (vitamin K2). DHNA synthetase converts O-succinyl-benzoyl-CoA (OSB-CoA) to 1,4-dihydroxy- 2-naphthoic acid (DHNA). - 4-chlorobenzoate dehalogenase (EC 3.8.1.6) [6], a *Pseudomonas* enzyme which catalyzes the conversion of 4-chlorobenzoate-CoA to 4-hydroxybenzoate-CoA. - A *Rhodobacter capsulatus* protein of unknown function (ORF257) [7]. - *Bacillus subtilis* putative polyketide biosynthesis proteins *pksH* and *pksI*. - *Escherichia coli* carnitine racemase (gene *caiD*) [8]. - *Escherichia coli* hypothetical protein *ygfG*. - Yeast hypothetical protein YDR036c. As a signature pattern for these enzymes, a conserved region rich in glycine and hydrophobic residues was selected.

Consensus pattern: [LIVM]-[STA]-x-[LIVM]-[DENQRHSTA]-G-x(3)-[AG](3)-x(4)-[LIVMST]-x-[CSTA]-[DQHP]-[LIVMFY]-

- [1] Minami-Ishii N., Taketani S., Osumi T., Hashimoto T. Eur. J. Biochem. 185:73-78(1989).
- [2] Mueller-Newen G., Stoffel W. Biol. Chem. Hoppe-Seyler 372:613-624(1991).
- [3] Palosaari P.M., Hiltunen J.K. J. Biol. Chem. 265:2446-2449(1990).
- [4] Nakahigashi K., Inokuchi H. Nucleic Acids Res. 18:4937-4937(1990).
- [5] Driscoll J.R., Taber H.W. J. Bacteriol. 174:5063-5071(1992).
- [6] Babbitt P.C., Kenyon G.L., Matin B.M., Charest H., Sylvestre M., Scholten J.D., Chang K.-H., Liang P.-H., Dunaway-Mariano D. Biochemistry 31:5594-5604(1992).
- [7] Beckman D.L., Kranz R.G. Gene 107:171-172(1991).
- [8] Eichler K., Bourgis F., Buchet A., Kleber H.-P., Mandrand-Berthelot M.-A. Mol. Microbiol. 13:775-786(1994).

166. (EF1BD) Elongation factor 1 beta/beta'/delta chain signatures

Eukaryotic elongation factor 1 (EF-1) is responsible for the GTP-dependent binding of aminoacyl-tRNAs to the ribosomes [1]. EF-1 is composed of four subunits: the alpha chain which binds GTP and aminoacyl-tRNAs, the gamma chain that probably plays a role in anchoring the complex to other cellular components and the beta and delta (or beta') chains. The beta and delta chains are highly similar proteins that both stimulate the exchange of GDP

210

bound to the alpha chain for GTP [2]. The beta and delta chains are hydrophilic proteins of around 23 to 31 Kd. Their C-terminal part seems important for the nucleotide exchange activity, while the N-terminal section is probably involved in the interaction with the gamma chain. Two signature patterns for this family of proteins were developed. The first
5 corresponds to an acidic region in the central section; the second, to the C-terminal extremity of these proteins

Consensus pattern: [DE]-[DEG]-[DE](2)-[LIVMF]-D-L-F-G-

Consensus pattern: [IV]-Q-S-x-D-[LIVM]-x-A-[FWM]-[NQ]-K-[LIVM]-

10

[1] Riis B., Rattan I.S., Clark B.F.C., Merrick W.C. Trends Biochem. Sci. 15:420-424(1990).

[2] van Damme H.T.F., Amons R., Karssies R., Timmers C.J., Janssen G.M.C., Moeller W. Biochim. Biophys. Acta 1050:241-247(1990).

15

167. (EF1G_domain) Elongation factor 1 gamma, conserved domain

168. (EFG_C) Elongation factor G C-terminus

20

This family is always found associated with GTP_EFTU. This family includes the carboxyl terminal regions of Elongation factor G, elongation factor 2 and some tetracycline resistance proteins.

25

169. (EFP) Elongation factor P signature

Elongation factor P (EF-P) [1] is a prokaryotic protein translation factor required for efficient peptide bond synthesis on 70S ribosomes from fMet-tRNA^{fMet}. EF-P is a protein of 21 Kd. It is evolutionary related to yeiP, an hypothetical protein from Escherichia coli. As a signature pattern, a conserved region located in the C-terminal part of these proteins was
30 selected.

Consensus pattern: K-x-[AV]-x(4)-G-x(2)-[LIV]-x-V-P-x(2)-[LIV]-x(2)-G-

[1] Aoki H., Adams S.-L., Turner M.A., Ganoza M.C. Biochimie 79:7-11(1997).

170. (EF TS) Elongation factor Ts signatures

In prokaryotes elongation factor Ts (EF-Ts) is a component of the elongation cycle of protein biosynthesis. It associates with the EF-Tu.GDP complex and induces the exchange of GDP to GTP, it remains bound to the aminoacyl-tRNA.EF-Tu.GTP complex up to the GTP hydrolysis stage on the ribosome [1].EF-Ts is also a component of the chloroplast protein biosynthetic machinery and is encoded in the genome of some algal chloroplast [2]. It is also present in mitochondria [3]. As signature patterns for EF-Ts, two conserved regions located in the N-terminal part of the protein have been selected.

Consensus pattern: L-R-x(2)-T-[GSDNQ]-x-[GS]-[LIVMF]-x(0,1)-[DENKAC]-x-K-[KRNEQS]-A-L-

Consensus pattern: E-[LIVM]-[NV]-[SCV]-[QE]-T-D-F-V-[SA]-[KRN]-

[1] Bubunencko M.G., Kireeva M.L., Gudkov A.T. Biochimie 74:419-425(1992).

[2] Kostrzewa M., Zetsche K. Plant Mol. Biol. 23:67-76(1993).

[3] Xin H., Worlax V.L., Burkhart W.A., Spremulli L.L. J. Biol. Chem. 270:17243-17249(1995).

171. (EMP24_GP25L) emp24/gp25L/p24 family

Members of this family are implicated in bringing cargo forward from the ER and binding to coat proteins by their cytoplasmic domains. Number of members: 30

Paccaud JP, Thomas DY, Bergeron JJ, Nilsson T, J Cell Biol 1998;140:751-765.

172. ENV_polyprotein

ENV polyprotein (coat polyprotein)

Number of members: 224

173. (ERG4_ERG24) Ergosterol biosynthesis ERG4/ERG24 family signatures

Two fungal enzymes involved in ergosterol biosynthesis and which act by reducing double bonds in precursors of ergosterol have been shown to be evolutionary related [1]. These are C-14 sterol reductase (gene ERG24 in budding yeast and *erg3* in *Neurospora Crassa*) and C-24(28) sterol reductase (gene ERG4 in budding yeast and *sts1* in fission yeast). Their sequences are also highly related to that of chicken lamin B receptor, which is thought to anchor the lamina to the inner nuclear membrane. These proteins are highly hydrophobic and seem to contain seven or eight transmembrane regions. As signature patterns, two conserved regions were selected. The first one is apparently located in a loop between the fourth and fifth transmembrane regions and the second is in the C-terminal section.

Consensus pattern: G-x(2)-[LIVM]-[YH]-D-x-[FYW]-x-G-x(2)-L-N-P-R -

Consensus pattern: [LIVM](2)-H-R-x(2)-R-D-x(3)-C-x(2)-K-Y-G-

[1] Lai M.H., Bard M., Pierson C.A., Alexander J.F., Goebel M., Carter G.T., Kirsch D.R. Gene 140:41-49(1994).

174. (ERM) Ezrin/radixin/moesin family

This family of proteins contain a band 4.1 domain (Band 41), at their amino terminus. This family represents the rest of these proteins.

[1] Yonemura S, Hirao M, Doi Y, Takahashi N, Kondo T, Tsukita S, J Cell Biol 1998;140:885-895.

175. ER lumen protein retaining receptor signatures

Proteins that reside in the lumen of the endoplasmic reticulum (ER) contain a C-terminal tetrapeptide (generally K-D-E-L or H-D-E-L) that serves as a signal for their retrieval (retrograde transport) from subsequent compartments of the secretory pathway. The signal is recognized by a receptor molecule that is believed to cycle between the cis side of the Golgi apparatus and the ER [1]. This protein is known as the ER lumen protein retaining receptor or also as the 'KDEL receptor'. It has been characterized in a variety of species, including fungi (gene ERD2), plants, Plasmodium, Drosophila and mammals. In mammals two highly related

forms of the receptor are known. Structurally, the receptor is a protein of about 220 residues that seems to contain seven transmembrane regions [2]. The N-terminal part (3 residues) is oriented toward the lumen while the C-terminal tail (about 12 residues) is cytoplasmic. There are three luminal and three cytoplasmic loops. Two signature patterns for these receptors were developed. The first pattern corresponds to the C-terminal half of the first cytoplasmic loop as well as most of the second transmembrane domain. The second pattern is a perfectly conserved decapeptide that corresponds to the central part of the fifth transmembrane domain.

Consensus pattern: G-I-S-x-[KR]-x-Q-x-L-[FY]-x-[LIV](2)-F-x(2)-R-Y-
Consensus pattern: L-E-[SA]-V-A-I-[LM]-P-Q-L-

[1] Pelham H.R.B. Curr. Opin. Cell Biol. 3:585-591(1991).

[2] Townsley F.M., Wilson D.W., Pelham H.R.B. EMBO J. 12:2821-2829(1993).

176. (ETF_beta) Electron transfer flavoprotein beta-subunit signature

The electron transfer flavoprotein (ETF) [1,2] serves as a specific electron acceptor for various mitochondrial dehydrogenases. ETF transfers electrons to the main respiratory chain via ETF-ubiquinone oxidoreductase. ETF is a heterodimer that consist of an alpha and a beta subunit and which bind one molecule of FAD per dimer. A similar system also exists in some bacteria. The beta subunit of ETF is a protein of about 28 Kd which is structurally related to the bacterial nitrogen fixation protein fixA which could play a role in a redox process and feed electrons to ferredoxin. Other related proteins are: - Escherichia coli hypothetical protein ydiQ. - Escherichia coli hypothetical protein ygcR. As a signature pattern for these proteins, a conserved region which is located in the central section was selected.

Consensus pattern: [IVA]-x-[KR]-x(2)-[DE]-[GD]-[GDE]-x(1,2)-[EQ]-x-[LIV]- x(4)-P-x-[LIVM](2)-[TAC]-

[1] Finocchiaro G., Ikeda Y., Ito M., Tanaka K. Prog. Clin. Biol. Res. 321:637-652(1990).

[2] Tsai M.H., Saier M.H. Jr. Res. Microbiol. 146:397-404(1995).

177. Endonuclease III signatures

Escherichia coli endonuclease III (EC 4.2.99.18) (gene nth) [1] is a DNA repair enzyme that acts both as a DNA N-glycosylase, removing oxidized pyrimidines from DNA, and as an apurinic/apyrimidinic (AP) endonuclease, introducing a single-strand nick at the site from which the damaged base was removed. Endonuclease III is an iron-sulfur protein that binds a single 4Fe-4S cluster. The 4Fe-4S cluster does not seem to be important for catalytic activity, but is probably involved in the proper positioning of the enzyme along the DNA strand [2]. Endonuclease III is evolutionary related to the following proteins: - Fission yeast endonuclease III homolog (gene nth1) [3]. - Escherichia coli and related protein DNA repair protein mutY, which is an adenine glycosylase. MutY is a larger protein (350 amino acids) than endonuclease III (211 amino acids). - Micrococcus luteus ultraviolet N-glycosylase/AP lyase which initiates repair at cis-syn pyrimidine dimers. - ORF10 in plasmid pFV1 of the thermophilic archaeobacteria Methanobacterium thermoformicum [4]. Restriction methylase m.MthTI, which is encoded by this plasmid, generates 5-methylcytosine which is subject to deamination resulting in G-T mismatches. This protein could correct these mismatches. - Yeast hypothetical protein YAL015c. - Fission yeast hypothetical protein SpAC26A3.02. - Caenorhabditis elegans hypothetical protein R10E4.5. - Methanococcus jannaschii hypothetical protein MJ0613. The 4Fe-4S cluster is bound by four cysteines which are all located in a 17 amino acid region at the C-terminal end of endonuclease III. A similar region is also present in the central section of mutY and in the C-terminus of ORF10 and of the Micrococcus UV endonuclease. The 4Fe-4S cluster region does not exist in YAL015c. Two signature patterns for these proteins were developed: the first corresponds to the core of the iron-sulfur binding domain, the second corresponds to the best conserved region in the catalytic core of these enzymes.

Consensus pattern: C-x(3)-[KRS]-P-[KRAGL]-C-x(2)-C-x(5)-C [The four C's are 4Fe-4S ligands]-

Consensus pattern: [GST]-x-[LIVMF]-P-x(5)-[LIVMW]-x(2,3)-[LI]-[PAS]-G-V-[GA]-x(3)-[GAC]-x(3)-[LIVM]-x(2)-[SALV]-[LIVMFYW]-[GANK]-

[1] Kuo C.-F., McRee D., Fisher C.L., O'Handley S.F., Cunningham R.P., Tainer J.A. Science 258:434-440(1992).

[2] Thomson A.J. Curr. Biol. 3:173-174(1993).

[3] Roldan-Arjona T., Anselmino C., Lindahl T. Nucleic Acids. Res. 3307-3312(1996).

[4] Noelling J., van Eeden F.J.M., Eggen R.I.L., de Vos W.M. Nucleic Acids Res. 20:6501-6507(1992).

5

178. (Epimerase) NAD dependent epimerase/dehydratase family

This family of proteins utilize NAD as a cofactor. The proteins in this family use nucleotide-sugar substrates for a variety of chemical reactions.

10 [1] Thoden JB, Hegeman AD, Wesenberg G, Chapeau MC, Frey PA, Holden HM, Biochemistry 1997;36:6294-6304.

179. Exonuclease

This family includes a variety of exonuclease proteins, such as ribonuclease T and the epsilon subunit of DNA polymerase III.

[1] Koonin EV, Deutscher MP, Nucleic Acids Res 1993;21:2521-2522.

20

180. ENTH

ENTH domain

[1] Kay BK, Yamabhai M, Wendland B, Emr SD; Medline: 99156083, Identification of a novel domain shared by putative components of the endocytic and cytoskeletal machinery.

25 Protein Sci 1999;8:435-438.

The ENTH (Epsin N-terminal homology) domain is found in proteins involved in endocytosis and cytoskeletal machinery. The function of the ENTH domain is unknown.

30 Number of members: 29

181. (eIF-1A) Eukaryotic initiation factor 1A signature

216

Eukaryotic translation initiation factor 1A (eIF-1A) [1] (formerly known as eIF-4C) is a protein that seems to be required for maximal rate of protein biosynthesis. It enhances ribosome dissociation into subunits and stabilizes the binding of the initiator Met-tRNA to 40S ribosomal subunits. eIF-1A is a hydrophilic protein of about 15 to 17 Kd. Archaeobacteria also seem to possess a eIF-1A homolog. As a signature pattern, a conserved region in the central section of these proteins was selected.

Consensus pattern: [IM]-x-G-x-[GS]-[KRH]-x(4)-[CL]-x-D-G-x(2)-R-x(2)-[RH]-I-x-G

[1] Wei C.-L., Kainuma M., Hershey J.W.B. *J. Biol. Chem.* 270:22788-22794(1995).

182. (eIF-5A) Eukaryotic initiation factor 5A hypusine signature

Eukaryotic initiation factor 5A (eIF-5A) (formerly known as eIF-4D) [1,2] is a small protein whose precise role in the initiation of protein synthesis is not known. It appears to promote the formation of the first peptide bond. eIF-5A seems to be the only eukaryotic protein to contain an hypusine residue. Hypusine is derived from lysine by the post-translational addition of a butylamino group (from spermidine) to the epsilon-amino group of lysine. The hypusine group is essential to the function of eIF-5A. A hypusine-containing protein has been found in archaeobacteria such as *Sulfolobus acidocaldarius* or *Methanococcus jannaschii*; this protein is highly similar to eIF-5A and could play a similar role in protein biosynthesis. The signature developed for eIF-5A is centered around the hypusine residue.

Consensus pattern: [PT]-G-K-H-G-x-A-K [The first K is modified to hypusine]

[1] Park M.H., Wolff E.C., Folk J.E. *Biofactors* 4:95-104(1993).

[2] Schnier J., Schwelberger H.G., Smit-McBride Z., Kang H.A., Hershey J.W.B. *Mol. Cell. Biol.* 11:3105-3114(1991).

183. (efhand) S-100/ICaBP type calcium binding protein signature

S-100 are small dimeric acidic calcium and zinc-binding proteins [1] abundant in the brain. They have two different types of calcium-binding sites: a low affinity one with a special

structure and a 'normal' EF-hand type high affinity site. The vitamin-D dependent intestinal calcium-binding proteins (ICaBP or calbindin 9 Kd) also belong to this family of proteins, but it does not form dimers. In the past years the sequences of many new members of this family have been determined (for reviews see [2,3,4]); in most cases the function of these proteins is not yet known, although it is becoming clear that they are involved in cell growth and differentiation, cell cycle regulation and metabolic control. These proteins are: - Calcyclin (Prolactin receptor associated protein (PRA); clatropin; 2a9; 5B10; S100A6). - Calpactin I light chain (p10; p11; 42c; S100A10). - Calgranulin A (cystic fibrosis antigen (CFAg); MIF related protein 8 (MRP- 8); p8; S100A8). - Calgranulin B (MIF related protein 14 (MRP-14); p14; S100A9). - Calgranulin C. - Calgizzarin (S100C). - Placental calcium-binding protein (CAPL) (18a2; pL98; 42a; p9K; MTS1; metastatin; S100A4). - Protein S-100D (S100A5). - Protein S-100E (S100A3). - Protein S-100L (CAN19; S100A2). - Placental protein S-100P (S100E). - Psoriasin (S100A7). - Chemotactic cytokine CP-10 [5]. - Protein MRP-126 [6]. - Trichohyalin [7]. This is a large intermediate filament-associated protein that associates with keratin intermediate filaments (KIF); it contains a S- 100 type domain in its N-terminal extremity. A number of these proteins are known to bind calcium while others are not (p10 for example). Our EF-hand detecting pattern will fail to pick those proteins which have lost their calcium-binding properties. A pattern was developed which unambiguously picks up proteins belonging to this family. This pattern spans the region of the EF-hand high affinity site but makes no assumptions on the calcium-binding properties of this site.

Consensus pattern: [LIVMFYW](2)-x(2)-[LK]-D-x(3)-[DN]-x(3)-[DNSG]-[FY]-x- [ES]-[FYVC]-x(2)-[LIVMFS]-[LIVMF]

[1] Baudier J. (In) Calcium and Calcium Binding proteins, Gerday C., Bollis L., Giller R., Eds., pp102-113, Springer Verlag, Berlin, (1988).

[2] Moncrief N.D., Kretsinger R.H., Goodman M. J. Mol. Evol. 30:522-562(1990).

[3] Kligman D., Hilt D.C. Trends Biochem. Sci. 13:437-443(1988).

[4] Schaefer B.W., Wicki R., Engelkamp D., Mattei M.-G., Heizmann C.W. Genomics 25:638-643(1995).

[5] Lackmann M., Cornish C.J., Simpson R.J., Moritz R.L., Geczy C.L. J. Biol. Chem. 267:7499-7504(1992).

[6] Nakano T., Graf T. *Oncogene* 7:527-534(1992).

[7] Lee S.-C., Kim I.-G., Marekov L.N., O'Keefe E.J., Parry D.A.D., Steinert P.M., *J. Biol. Chem.* 268:12164-12176(1993).

5 EF-hand calcium-binding domain

Many calcium-binding proteins belong to the same evolutionary family and share a type of calcium-binding domain known as the EF-hand [1 to 5]. This type of domain consists of a twelve residue loop flanked on both side by a twelve residue alpha-helical domain. In an EF-hand loop the calcium ion is coordinated in a pentagonal bipyramidal configuration. The six residues involved in the binding are in positions 1, 3, 5, 7, 9 and 12; these residues are denoted by X, Y, Z, -Y, -X and -Z. The invariant Glu or Asp at position 12 provides two oxygens for liganding Ca (bidentate ligand).

Listed below are the proteins which are known to contain EF-hand regions. For each type of protein the total number of EF-hand regions known or supposed to exist is indicated between parenthesis. This number does not include regions which clearly have lost their calcium-binding properties, or the atypical low-affinity site (which spans thirteen residues) found in the S-100/ICaBP family of proteins [6].

- Aequorin and Renilla luciferin binding protein (LBP) (Ca=3).
- Alpha actinin (Ca=2). - Calbindin (Ca=4).
- Calcineurin B subunit (protein phosphatase 2B regulatory subunit) (Ca=4).
- Calcium-binding protein from *Streptomyces erythraeus* (Ca=3?).
- Calcium-binding protein from *Schistosoma mansoni* (Ca=2?).
- Calcium-binding proteins TCBP-23 and TCBP-25 from *Tetrahymena thermophila* (Ca=4?). - Calcium-dependent protein kinases (CDPK) from plants (Ca=4).
- Calcium vector protein from amphoxius (Ca=2).
- Calcyphosin (thyroid protein p24) (Ca=4?).
- Calmodulin (Ca=4, except in yeast where Ca=3).
- Calpain small and large chains (Ca=2). - Calretinin (Ca=6).
- Calcyclin (prolactin receptor associated protein) (Ca=2).
- Caltractin (centrin) (Ca=2 or 4).
- Cell Division Control protein 31 (gene CDC31) from yeast (Ca=2?).

- Diacylglycerol kinase (EC 2.7.1.107) (DGK) (Ca=2).
- FAD-dependent glycerol-3-phosphate dehydrogenase (EC 1.1.99.5) from mammals (Ca=1). - Fimbrin (plastin) (Ca=2).
- Flagellar calcium-binding protein (1f8) from Trypanosoma cruzi (Ca=1 or 2).
- 5 - Guanylate cyclase activating protein (GCAP) (Ca=3).
- Inositol phospholipid-specific phospholipase C isozymes gamma-1 and delta-1 (Ca=2) [10]. - Intestinal calcium-binding protein (ICaBPs) (Ca=2).
- MIF related proteins 8 (MRP-8 or CFAG) and 14 (MRP-14) (Ca=2).
- Myosin regulatory light chains (Ca=1). - Oncomodulin (Ca=2).
- 10 - Osteonectin (basement membrane protein BM-40) (SPARC) and proteins that contains an 'osteonectin' domain (QR1, matrix glycoprotein SC1) (see the entry <PDOC00535>) (Ca=1). - Parvalbumins alpha and beta (Ca=2).
- Placental calcium-binding protein (18a2) (nerve growth factor induced protein 42a) (p9k) (Ca=2).
- 15 - Recoverins (visinin, hippocalcin, neurocalcin, S-modulin) (Ca=2 to 3).
- Reticulocalbin (Ca=4). - S-100 protein, alpha and beta chains (Ca=2).
- Sarcoplasmic calcium-binding protein (SCPs) (Ca=2 to 3).
- Sea urchin proteins Spec 1 (Ca=4), Spec 2 (Ca=4?), Lps-1 (Ca=8).
- Serine/threonine protein phosphatase rdgc (EC 3.1.3.16) from Drosophila (Ca=2) - Sorcin V19 from hamster (Ca=2). - Spectrin alpha chain (Ca=2).
- 20 - Squidulin (optic lobe calcium-binding protein) from squid (Ca=4).
- Troponins C; from skeletal muscle (Ca=4), from cardiac muscle (Ca=3), from arthropods and molluscs (Ca=2).

There has been a number of attempts [7,8] to develop patterns that pick-up EF-hand regions, but these studies were made a few years ago when not so many different families of calcium-binding proteins were known. Therefore a new pattern was developed which takes into account all published sequences. This pattern includes the complete EF-hand loop as well as the first residue which follows the loop and which seem to always be hydrophobic.

30 -Consensus pattern: D-x-[DNS]-{ILVFYW}-{DENSTG}-[DNQGHRK]-{GP}-{LIVMC}-[DENQSTAGC]-x(2)-[DE]-[LIVMFYW]
 -Note: positions 1 (X), 3 (Y) and 12 (-Z) are the most conserved.

-Note: the 6th residue in an EF-hand loop is, in most cases a Gly, but the number of exceptions to this 'rule' has gradually increased and therefore the pattern should include all the different residues which have been shown to exist in this position in functional Ca-binding sites.

- 5 -Note: the pattern will, in some cases, miss one of the EF-hand regions in some proteins with multiple EF-hand domains.

[1] Kawasaki H., Kretsinger R.H. Protein Prof. 2:305-490(1995).[2] Kretsinger R.H. Cold Spring Harbor Symp. Quant. Biol. 52:499-510(1987).

10 [3] Moncrief N.D., Kretsinger R.H., Goodman M. J. Mol. Evol. 30:522-562(1990).

[4] Nakayama S., Moncrief N.D., Kretsinger R.H. J. Mol. Evol. 34:416-448(1992).

[5] Heizmann C.W., Hunziker W. Trends Biochem. Sci. 16:98-103(1991).

[6] Kligman D., Hilt D.C. Trends Biochem. Sci. 13:437-443(1988).

[7] Strynadka N.C.J., James M.N.G.

15 Annu. Rev. Biochem. 58:951-98(1989).

[8] Haiech J., Sallantin J. Biochimie 67:555-560(1985).

[9] Chauvaux S., Beguin P., Aubert J.-P., Bhat K.M., Gow L.A., Wood T.M., Bairoch A. Biochem. J. 265:261-265(1990).

[10] Bairoch A., Cox J.A. FEBS Lett. 269:454-456(1990).

184. Enolase signature

Enolase (EC 4.2.1.11) is a glycolytic enzyme that catalyzes the dehydration of 2-phospho-D-glycerate to phosphoenolpyruvate [1]. It is a dimeric enzyme that requires magnesium both for catalysis and stabilizing the dimer. Enolase is probably found in all organisms that metabolize sugars. In vertebrates, there are three different tissue-specific isozymes: alpha present in most tissues, beta in muscles and gamma found only in nervous tissues. Tau-crystallin, one of the major lens proteins in some fish, reptiles and birds, has been shown [2] to be evolutionary related to enolase. As a signature pattern for enolase, the best conserved region was selected, it is located in the C-terminal third of the sequence.-

Consensus pattern: [LIV](3)-K-x-N-Q-I-G-[ST]-[LIV]-[ST]-[DE]-[STA]

[1] Lebioda L., Stec B., Brewer J.M. J. Biol. Chem. 264:3685-3693(1989).

[2] Wistow G., Piattigorsky J. Science 236:1554-1556(1987).

185. (F-actin_cap_A) F-actin capping protein alpha subunit signatures

5 The F-actin capping protein binds in a calcium-independent manner to the fast growing ends of actin filaments (barbed end) thereby blocking the exchange of subunits at these ends. Unlike gelsolin and severin this protein does not sever actin filaments. The F-actin capping protein is a heterodimer composed of two unrelated subunits: alpha and beta. The alpha subunit is a protein of about 268 to 286 amino acid residues whose sequence is well
10 conserved in eukaryotic species [1]. As signature patterns two highly conserved regions in the C-terminal section of the alpha subunit were selected.

Consensus pattern: V-H-[FY](2)-E-D-G-N-V

Consensus pattern: F-K-[AE]-L-R-R-x-L-P-

[1] Cooper J.A., Caldwell J.E., Gattermeir D.J., Torres M.A., Amatruda J.F., Casella J.F. Cell Motil. Cytoskeleton 18:204-214(1991).

186. F-box domain

[1] Bai C, Sen P, Hofmann K, Ma L, Goebel M, Harper JW, Elledge SJ, Cell 1996;86:263-274. [2] Skowyra D, Craig KL, Tyers M, Elledge SJ, Harper JW, Cell 1997;91:209-219.

187. F-protein

Negative factor, (F-Protein) or Nef.

[1] Arold S, Franken P, Strub M-P, Hoh F, Benichou S, Benarous R, Dumas C; Medline: 98035457, The crystal structure of HIV-1 Nef protein bound to the Fyn kinase SH3 domain suggests a role for this complex in altered T cell receptor signalling Structure 1997;5:1361-1372.

Nef protein accelerates virulent progression of AIDS by its interaction with cellular proteins involved in signal transduction and host cell activation. Nef has been shown to bind specifically to a subset of the Src kinase family.

5 Number of members: 1013

188. (FAD_binding_2)

Fumarate reductase / succinate dehydrogenase FAD-binding site

10

In bacteria two distinct, membrane-bound, enzyme complexes are responsible for the interconversion of fumarate and succinate (EC 1.3.99.1): fumarate reductase (Frd) is used in anaerobic growth, and succinate dehydrogenase (Sdh) is used in aerobic growth. Both complexes consist of two main components: a membrane-extrinsic component composed of a FAD-binding flavoprotein and an iron-sulfur protein; and an hydrophobic component composed of a membrane anchor protein and/or a cytochrome B.

In eukaryotes mitochondrial succinate dehydrogenase (ubiquinone) (EC 1.3.5.1) is an enzyme composed of two subunits: a FAD flavoprotein and and iron-sulfur protein.

20

The flavoprotein subunit is a protein of about 60 to 70 Kd to which FAD is covalently bound to a histidine residue which is located in the N-terminal section of the protein [1]. The sequence around that histidine is well conserved in Frd and Sdh from various bacterial and eukaryotic species [2] and can be used as a signature pattern.

25

Consensus patternR-[ST]-H-[ST]-x(2)-A-x-G-G [H is the FAD binding site] Sequences known to belong to this class detected by the pattern ALL.

30

[1] Blaut M., Whittaker K., Valdovinos A., Ackrell B.A., Gunsalus R.P., Cecchini G. J. Biol. Chem. 264:13599-13604(1989).

[2] Birch-Machin M.A., Farnsworth L., Ackrell B.A., Cochran B., Jackson S., Bindoff L.A., Aitken A., Diamond A.G., Turnbull D.M. J. Biol. Chem. 267:11553-11558(1992).

189. Fatty acid desaturases signatures (FA_desaturase)

Fatty acid desaturases (EC 1.14.99.-) are enzymes that catalyze the insertion of a double bond at the delta position of fatty acids. There seems to be two distinct families of fatty acid

5 desaturases which do not seem to be evolutionary related. Family 1 is composed of: - Stearoyl-CoA desaturase (SCD) (EC 1.14.99.5) [1]. SCD is a key regulatory enzyme of unsaturated fatty acid biosynthesis. SCD introduces a cis double bond at the delta(9) position of fatty acyl-CoA's such as palmitoleoyl- and oleoyl-CoA. SCD is a membrane-bound enzyme that is thought to function as a part of a multienzyme complex in the endoplasmic
10 reticulum of vertebrates and fungi. As a signature pattern for this family a conserved region in the C-terminal part of these enzymes was selected, this region is rich in histidine residues and in aromatic residues. Family 2 is composed of: - Plants stearoyl-acyl-carrier-protein desaturase (EC 1.14.99.6) [2], these enzymes catalyze the introduction of a double bond at the delta(9) position of stearoyl-ACP to produce oleoyl-ACP. This enzyme is responsible for
15 the conversion of saturated fatty acids to unsaturated fatty acids in the synthesis of vegetable oils. - Cyanobacteria desA [3] an enzyme that can introduce a second cis double bond at the delta(12) position of fatty acid bound to membranes glycerolipids. DesA is involved in chilling tolerance; the phase transition temperature of lipids of cellular membranes being dependent on the degree of unsaturation of fatty acids of the membrane lipids. As a signature
20 pattern for this family a conserved region in the C-terminal part of these enzymes was selected.

Consensus pattern: G-E-x-[FY]-H-N-[FY]-H-H-x-F-P-x-D-Y-

Consensus pattern: [ST]-[SA]-x(3)-[QR]-[LI]-x(5,6)-D-Y-x(2)-[LIVMFYW]-[LIVM]- [DE]-

[1] Kaestner K.H., Ntambi J.M., Kelly T.J. Jr., Lane M.D. J. Biol. Chem. 264:14755-14761(1989).

[2] Shanklin J., Somerville C.R. Proc. Natl. Acad. Sci. U.S.A. 88:2510-2514(1991).

[3] Wada H., Gombos Z., Murata N. Nature 347:200-203(1990).

190. Fructose-1-6-bisphosphatase active site (FBPase)

Fructose-1,6-bisphosphatase (EC 3.1.3.11) (FBPase) [1], a regulatory enzyme in gluconeogenesis, catalyzes the hydrolysis of fructose 1,6-bisphosphate to fructose 6-phosphate. It is involved in many different metabolic pathways and found in most organisms. Sedoheptulose-1,7-bisphosphatase (EC 3.1.3.37) (SBPase) [2] is an enzyme found plant chloroplast and in photosynthetic bacteria that catalyzes the hydrolysis of sedoheptulose 1,7-bisphosphate to sedoheptulose 7-phosphate, a step in the Calvin's reductive pentose phosphate cycle. It is functionally and structurally related to FBPase. In mammalian FBPase, a lysine residue has been shown to be involved in the catalytic mechanism [3]. The region around this residue is highly conserved and can be used as a signature pattern for FBPase and SBPase. It must be noted that, in some bacterial FBPase sequences, the active site lysine is replaced by an arginine

Consensus pattern: [AG]-[RK]-L-x(1,2)-[LIV]-[FY]-E-x(2)-P-[LIVM]-[GSA] [K/R is the active site residue]-

[1] Benkovic S.J., DeMaine M.M. Adv. Enzymol. 53:45-82(1982).

[2] Raines C.A., Lloyd J.C., Willingham N.M., Potts S., Dyer T.A. Eur. J. Biochem. 205:1053-1059(1992).

[3] Ke H., Thorpe C.M., Seaton B.A., Lipscomb W.N., Marcus F. J. Mol. Biol. 212:513-539(1989).

191. FGGY family of carbohydrate kinases signatures *

It has been shown [1] that four different type of carbohydrate kinases seem to be evolutionary related. These enzymes are: - L-fucolokinase (EC 2.7.1.51) (gene fucK). - Gluconokinase (EC 2.7.1.12) (gene gntK). - Glycerokinase (EC 2.7.1.30) (gene glpK). - Xylulokinase (EC 2.7.1.17) (gene xylB). - L-xylulose kinase (EC 2.7.1.53) (gene lyxK). These enzymes are proteins of from 480 to 520 amino acid residues. As consensus patterns for this family of kinases two conserved regions were selected, one in the central section, the other in the C-terminal section.

Consensus pattern: [MFYGS]-x-[PST]-x(2)-K-[LIVMFYW]-x-W-[LIVMF]-x-[DENQTKR]-[ENQH]-

225

Consensus pattern: [GSA]-x-[LIVMFYW]-x-G-[LIVM]-x(7,8)-[HDENQ]-[LIVMF]-x(2)-
[AS]-[STAI VM]-[LIVMFY]-[DEQ]-

[1] Reizer A., Deutscher J., Saier M.H. Jr., Reizer J. Mol. Microbiol. 5:1081-1089(1991).

5

192. FKBP-type peptidyl-prolyl cis-trans isomerase signatures/profile (FKBP)

FKBP [1,2,3] is the major high-affinity binding protein, in vertebrates, for the immunosuppressive drug FK506. It exhibits peptidyl-prolyl cis-trans isomerase activity (EC 5.2.1.8) (PPIase or rotamase). PPIase is an enzyme that accelerates protein folding by catalyzing the cis-trans isomerization of proline imidic peptide bonds in oligopeptides [4]. At least three different forms of FKBP are known in mammalian species: - FKBP-12, which is cytosolic and inhibited by both FK506 and rapamycin. - FKBP-13, which is membrane associated and inhibited by both FK506 and rapamycin. - FKBP-25, which is preferentially inhibited by rapamycin. These forms of FKBP are evolutionary related and show extensive similarities[5,6,7] with the following proteins: - Fungal FKBP. - Mammalian hsp binding immunophilin (HBI) (also called p59). HBI is a protein which binds to hsp90 and contains two FKBP-like domains in its N- terminal section - the first of which seems to be functional. - The C-terminal part of the cell-surface protein mip from Legionella; a protein associated with macrophage infection by an unknown mechanism. - Escherichia coli slyD [8], a protein with a N-terminal FKBP domain followed by an histidine-rich metal-binding domain. - Escherichia coli fkpA. - Escherichia coli fklB (FKBP22). - Escherichia coli slpA. - Bacterial trigger factor (Tig). - Streptomyces hygroscopus and chrysomallus FK506-binding protein. - Chlamydia trachomatis 27 Kd membrane protein. - Neisseria meningitidis strain C114 PPIase. - Probable PPIases from Haemophilus influenzae (HI0754), Methanococcus jannaschii (MJ0278 and MJ0825), Pseudomonas fluorescens and Pseudomonas aeruginosa. Two signature patterns for these proteins were developed. One is based on a conserved region in the N-terminus of FKBP, the other is located in the central section. The profile for FKBP spans the complete domain.

30

Consensus pattern: [LIVMC]-x-[YF]-x-[GVL]-x(1,2)-[LFT]-x(2)-G-x(3)-[DE]- [STAEQK]-
[STAN]-

Consensus pattern: [LIVMFY]-x(2)-[GA]-x(3,4)-[LIVMF]-x(2)-[LIVMFHK]-x(2)-G- x(4)-
[LIVMF]-x(3)-[PSGAQ]-x(2)-[AG]-[FY]-G--

[1] Tropschug M., Wachter E., Mayer S., Schoenbrunner E.R., Schmid F.X. Nature 346:674-
677(1990).

[2] Stein R.L. Curr. Biol. 1:234-236(1991).

[3] Siekierka J.J., Widerrecht G., Greulich H., Boulton D., Hung S.H.Y., Cryan J., Hodges
P.J., Sigal N.H. J. Biol. Chem. 265:21011-21015(1990).

[4] Fischer G., Schmid F.X. Biochemistry 29:2205-2212(1990).

[5] Trandinh C.C., Pao G.M., Saier M.H. Jr. FASEB J. 6:3410-3420(1992).

[6] Galat A. Eur. J. Biochem. 216:689-707(1993).

[7] Hacker J., Fischer G. Mol. Microbiol. 10:445-456(1993).

[8] Wuelfing C., Lomardero J., Plueckthun A. J. Biol. Chem. 269:2895-2901(1994).

193. MAPEG family (aka: FLAP/GST2/LTC4S family signature)

The following mammalian proteins are evolutionary related [1]:

- Leukotriene C4 synthase (EC 2.5.1.37) (gene LTC4S), an enzyme that catalyzes
the production of LTC4 from LTA4.
- Microsomal glutathione S-transferase II (EC 2.5.1.18) (GST-II) (gene GST2), an
enzyme that can also produces LTC4 from LTA4.
- 5-lipoxygenase activating protein (gene FLAP), a protein that seems to be
required for the activation of 5-lipoxygenase.

These are proteins of 150 to 160 residues that contain three transmembrane segments.

As a signature pattern, a conserved region between the first and second transmembrane
domains was selected.

Consensus pattern: G-x(3)-F-E-R-V-[FY]-x-A-[NQ]-x-N-C

[1] Jakobsson P.-J., Mancini J.A., Ford-Hutchinson A.W. J. Biol. Chem. 271:22203-
22210(1996).

194. FMN-dependent alpha-hydroxy acid dehydrogenases active site (FMN_dh)

A number of oxidoreductases that act on alpha-hydroxy acids and which are FMN-containing flavoproteins have been shown [1,2,3] to be structurally related; these enzymes are: - Lactate dehydrogenase (EC 1.1.2.3), which consists of a dehydrogenase domain and a heme-binding domain called cytochrome b2 and which catalyzes the conversion of lactate into pyruvate. -

5 Glycolate oxidase (EC 1.1.3.15) ((S)-2-hydroxy-acid oxidase), a peroxisomal enzyme that catalyzes the conversion of glycolate and oxygen to glyoxylate and hydrogen peroxide. -

Long chain alpha-hydroxy acid oxidase from rat (EC 1.1.3.15), a peroxisomal enzyme. -

Lactate 2-monooxygenase (EC 1.13.12.4) (lactate oxidase) from *Mycobacterium smegmatis*, which catalyzes the conversion of lactate and oxygen to acetate, carbon dioxide and water. -

10 (S)-mandelate dehydrogenase from *Pseudomonas putida* (gene *mdlB*), which catalyzes the reduction of (S)-mandelate to benzoylformate. The first step in the reaction mechanism of these enzymes is the abstraction of the proton from the alpha-carbon of the substrate producing a carbanion which can subsequently attach to the N5 atom of FMN. A conserved histidine has been shown [4] to be involved in the removal of the proton. The region around this active site residue is highly conserved and contains an arginine residue which is involved in substrate binding.

Consensus pattern: S-N-H-G-[AG]-R-Q [H is the active site residue] [R is a substrate-binding residue]-

[1] Giegel D.A., Williams C.H. Jr., Massey V. J. Biol. Chem. 265:6626-6632(1990).

[2] Tsou A.Y., Ransom S.C., Gerlt J.A., Buechter D.D., Babbitt P.C., Kenyon G.L. Biochemistry 29:9856-9862(1990).

[3] Le K.H.D., Lederer F. J. Biol. Chem. 266:20877-20880(1991).

25 [4] Lindqvist Y., Branden C.-I. J. Biol. Chem. 264:3624-3628(1989).

195. Flavin-binding monooxygenase-like (FMO-like)

This family includes FMO proteins, cyclohexanone monooxygenase

196. (FPGS)

Folylpolyglutamate synthase signatures (aka *Mur_ligase*)

Folylpolyglutamate synthase (EC 6.3.2.17) (FPGS) [1] is the enzyme of folate metabolism that catalyzes ATP-dependent addition of glutamate moieties to tetrahydrofolate.

- 5 Its sequence is moderately conserved between prokaryotes (gene folC) and eukaryotes. We developed two signature patterns based on the conserved regions which are rich in glycine residues and could play a role in the catalytical activity and/or in substrate binding.

10 Consensus pattern [LIVMFY]-x-[LIVM]-[STAG]-G-T-[NK]-G-K-x-[ST]-x(7)-[LIVM](2)-x(3)-[GSK] Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern[LIVMFY](2)-E-x-G-[LIVM]-[GA]-G-x(2)-D-x-[GST]-x-[LIVM](2)
Sequences known to belong to this class detected by the pattern ALL.

15 [1] Shane B., Garrow T., Brenner A., Chen L., Choi Y.J., Hsu J.C., Stover P. Adv. Exp. Med. Biol. 338:629-634(1993).

□

20 197. FYVE zinc finger

The FYVE zinc finger is named after four proteins that it has been found in: Fab1, YOTB/ZK632.12, Vac1, and EEA1. The FYVE finger has been shown to bind two Zn⁺⁺ ions [1]. The FYVE finger has eight potential zinc coordinating cysteine positions. Many members of this family also include two histidines in a motif R+HHC+XCG, where +
25 represents a charged residue and X any residue. Members were included which do not conserve these histidine residues but are clearly related.

[1] Stenmark H, Aasland R, Toh BH, D'Arrigo A, J Biol Chem 1996;271:24048-24054. [2] Gaullier JM, Simonsen A, D'Arrigo A, Bremnes B, Stenmark H, Aasland R, Nature 1998;394:432-433.

30

198. F_actin_cap_B

F-actin capping protein beta subunit signature

The F-actin capping protein binds in a calcium-independent manner to the fast growing ends of actin filaments (barbed end) thereby blocking the exchange of subunits at these ends.

Unlike gelsolin and severin this protein does not sever actin filaments. The F-actin capping protein is a heterodimer composed of two unrelated subunits: alpha and beta.

The beta subunit is a protein of about 280 amino acid residues whose sequence is well conserved in eukaryotic species [1]. As a signature pattern a conserved hexapeptide in the N-terminal section of the beta subunit was selected.

Consensus pattern: C-D-Y-N-R-D Sequences known to belong to this class detected by the pattern ALL.

[1] Amatruda J.F., Cannon J.F., Tatchell K., Hug C., Cooper J.A. Nature 344:352-354(1990).

199. Isopenicillin N synthetase signatures (Fe_Asc_oxidored)

Isopenicillin N synthetase (IPNS) [1,2] is a key enzyme in the biosynthesis of penicillin and cephalosporin. In the presence of oxygen, it removes iron and ascorbate, four hydrogen atoms from L-(alpha-aminoadipyl)-L-cysteinyl-d-valine to form the azetidinone and thiazolidine rings of isopenicillin. IPNS is an enzyme of about 330 amino-acid residues. Two cysteines are conserved in fungal and bacterial IPNS sequences; these may be involved in iron-binding and/or substrate-binding. Cephalosporium acremonium DAOCS/DACS [3] is a bifunctional enzyme involved in cephalosporin biosynthesis. The DAOCS domain, which is structurally related to IPNS, catalyzes the step from penicillin N to deacetoxy-cephalosporin C - used as a substrate by DACS to form deacetylcephalosporin C. Streptomyces clavuligerus possesses a monofunctional DAOCS enzyme (gene cefE) [4] also related to IPNS. Two signature patterns for these enzymes were derived, centered around the conserved cysteine residues.

Consensus pattern: [RK]-x-[STA]-x(2)-S-x-C-Y-[SL]-

Consensus pattern: [LIVM](2)-x-C-G-[STA]-x(2)-[STAG]-x(2)-T-x-[DNG]-

[1] Martin J.F. Trends Biotechnol. 5:306-308(1987).

- [2] Chen G., Shiffman D., Mevarech M., Aharonowitz Y. Trends Biotechnol. 8:105-111(1990).
- [3] Samson S.M., Dotzla J.E., Slisz M.L., Becker G.W., van Frank R.M., Veal L.E., Yeh W.K., Miller J.R., Queener S.W., Ingolia T.D. Bio/Technology 5:1207-1214(1987).
- [4] Kovacevic S., Weigel B.J., Tobin M.B., Ingolia T.D., Miller J.R. J. Bacteriol. 171:754-760(1989).

200. Fibrillarin signature

Fibrillarin [1] is a component of a nucleolar small nuclear ribonucleoprotein(SnRNP) particle thought to participate in the first step of the processing of pre-rRNA. In mammals, fibrillarin is associated with the U3, U8 and U13small nuclear RNAs [2]. Fibrillarin is an extremely well conserved protein of about 320 amino acid residues. Structurally it consists of three different domains: - An N-terminal domain of about 80 amino acids which is very rich in glycine and contains a number of dimethylated arginine residues (DMA). - A central domain of about 90 residues which resembles that of RNA-binding proteins and contains an octameric sequence similar to the RNP-2 consensus found in such proteins. - A C-terminal alpha-helical domain. A protein evolutionary related to fibrillarin has been found [3] in archaeobacteria such as Methanococcus vanniellii or voltae. This protein (gene flpA) is involved in pre-rRNA processing. It lacks the Gly/Arg-rich N-terminal domain. As a signature pattern, a region was selected that starts with and encompasses the RNP-2 like octapeptide sequence.

Consensus pattern: [GST]-[LIVMAP]-V-Y-A-[IV]-E-[FY]-[SA]-x-R-x(2)-R-[DE] -

- [1] Aris J.P., Blobel G. Proc. Natl. Acad. Sci. U.S.A. 88:931-935(1991).
- [2] Bandziulis R.J., Swanson M.S., Dreyfuss G. Genes Dev. 3:431-437(1989).
- [3] Agha-Amiri K. J. Bacteriol. 176:2124-2127(1994).

201. Filamin/ABP280 repeat

[1] Fucini P, Renner C, Herberhold C, Noegel AA, Holak TA, Nat Struct Biol 1997;4:223-230.

202. Fucosyl transferase

This family of Fucosyltransferases are the enzymes transferring
 5 fucose from GDP-Fucose to GlcNAc in an alpha1,3 linkage [1].

[1] Breton C, Oriol R, Imberty A; Glycobiology 1998;8:87-94.

203. 2Fe-2S ferredoxins, iron-sulfur binding region signature (fer2A)

10 Ferredoxins [1] are a group of iron-sulfur proteins which mediate electron transfer in a wide variety of metabolic reactions. Ferredoxins can be divided into several subgroups depending upon the physiological nature of the iron sulfur cluster(s) and according to sequence similarities. One of these subgroups are the 2Fe-2S ferredoxins, which are proteins or domains of around one hundred amino acid residues that bind a single 2Fe-2S iron-sulfur
 15 cluster. The proteins that are known [2] to belong to this family are listed below. - Ferredoxin from photosynthetic organisms; namely plants and algae where it is located in the chloroplast or cyanelle; and cyanobacteria. - Ferredoxin from archaebacteria of the Halobacterium genus. - Ferredoxin IV (gene pftA) and V (gene fdxD) from Rhodobacter capsulatus. - Ferredoxin in the toluene degradation operon (gene xylT) and naphthalene degradation operon (gene nahT)
 20 of Pseudomonas putida. - Hypothetical Escherichia coli protein yfaE. - The N-terminal domain of the bifunctional ferredoxin/ferredoxin reductase electron transfer component of the benzoate 1,2-dioxygenase complex (gene benC) from Acinetobacter calcoaceticus, the toluene 4-monooxygenase complex (gene tmoF), the toluate 1,2-dioxygenase system (gene xylZ), and the xylene monooxygenase system (gene xylA) from Pseudomonas. - The N-
 25 terminal domain of phenol hydroxylase protein p5 (gene dmpP) from Pseudomonas Putida. - The N-terminal domain of methane monooxygenase component C (gene mmoC) from Methylococcus capsulatus . - The C-terminal domain of the vanillate degradation pathway protein vanB in a Pseudomonas species. - The N-terminal domain of bacterial fumarate reductase iron-sulfur protein (gene frdB). - The N-terminal domain of CDP-6-deoxy-3,4-
 30 glucoseen reductase (gene ascD) from Yersinia pseudotuberculosis. - The central domain of eukaryotic succinate dehydrogenase (ubiquinone) iron- sulfur protein. - The N-terminal domain of eukaryotic xanthine dehydrogenase. - The N-terminal domain of eukaryotic aldehyde oxidase. In the 2Fe-2S ferredoxins, four cysteine residues bind the iron-sulfur

cluster. Three of these cysteines are clustered together in the same region of the protein. Our signature pattern spans that iron-sulfur binding region.

Consensus pattern: C-{C}-{C}-[GA]-{C}-C-[GAST]-{CPDEKRHFYW}-C [The three C's are 2Fe-2S ligands]-

[1] Meyer J. Trends Ecol. Evol. 3:222-226(1988).[2] Harayama S., Polissi A., Rekik M. FEBS Lett. 285:85-88(1991).

Adrenodoxin family, iron-sulfur binding region signature (fer2B)

Ferredoxins [1] are a group of iron-sulfur proteins which mediate electron transfer in a wide variety of metabolic reactions. Ferredoxins can be divided into several subgroups depending upon the physiological nature of the iron sulfur cluster(s) and according to sequence similarities. One family of ferredoxins groups together the following proteins that all bind a single 2Fe-2S iron-sulfur cluster: - Adrenodoxin (ADX) (adrenal ferredoxin), a vertebrate mitochondrial protein which transfers electrons from adrenodoxin reductase to cytochrome P450_{scc}, which is involved in cholesterol side chain cleavage. - Putidaredoxin (PTX), a *Pseudomonas putida* protein which transfers electrons from putidaredoxin reductase to cytochrome P450_{cam}, which is involved in the oxidation of camphor. - Terpredoxin [2], a *Pseudomonas* protein which transfers electrons from terpredoxin reductase to cytochrome P450_{terp}, which is involved in the oxidation of alpha-terpineol. - Rhodocoxin [3], a *Rhodococcus* protein which transfers electrons from rhodocoxin reductase to cytochrome CYP116 (thcB), which is involved in the degradation of thiocarbamate herbicides. - *Escherichia coli* ferredoxin (gene fdx) [4] whose exact function is not yet known. - *Rhodobacter capsulatus* ferredoxin VI [5], which may transfer electrons to a yet uncharacterized oxygenase. - *Caulobacter crescentus* ferredoxin (gene fdxB) [6]. In these proteins, four cysteine residues bind the iron-sulfur cluster. Three of these cysteines are clustered together in the same region of the protein. Our signature pattern spans that iron-sulfur binding region.

Consensus pattern: C-x(2)-[STAQ]-x-[STAMV]-C-[STA]-T-C-[HR] [The three C's are 2Fe-2S ligands]-

- [1] Meyer J. Trends Ecol. Evol. 3:222-226(1988).
- [2] Peterson J.A., Lu J.-Y., Geisselsoder J., Graham-Lorence S., Carmona C., Witney F., Lorence M.C. J. Biol. Chem. 267:14193-14203(1992).
- [3] Nagy I., Schoofs G., Compennolle F., Proost P., Vanderleyden J., De Mot R. J. Bacteriol. 177:676-687(1995).
- [4] Ta D.T., Vickery L.E. J. Biol. Chem. 267:11120-11125(1992).
- [5] Naud I., Vincon M., Garin J., Gaillard J., Forest E., Jouanneau Y. Eur. J. Biochem. 222:933-939(1994).
- [6] Amemiya K EMBL/Genbank: X51607.

204. 4Fe-4S ferredoxins, iron-sulfur binding region signature (fer4)

Ferredoxins [1] are a group of iron-sulfur proteins which mediate electron transfer in a wide variety of metabolic reactions. Ferredoxins can be divided into several subgroups depending upon the physiological nature of the iron-sulfur cluster(s). One of these subgroups are the 4Fe-4S ferredoxins, which are found in bacteria and which are thus often referred as 'bacterial-type' ferredoxins. The structure of these proteins [2] consists of the duplication of a domain of twenty six amino acid residues; each of these domains contains four cysteine residues that bind to a 4Fe-4S center. A number of proteins have been found [3] that include one or more 4Fe-4S binding domains similar to those of bacterial-type ferredoxins. These proteins are listed below (references are only provided for recently determined sequences). -

The iron-sulfur proteins of the succinate dehydrogenase and the fumarate reductase complexes (EC 1.3.99.1). These enzyme complexes, which are components of the tricarboxylic acid cycle, each contain three subunits: a flavoprotein, an iron-sulfur protein, and a b-type cytochrome. The iron-sulfur proteins contain three different iron-sulfur centers: a 2Fe-2S, a 3Fe-3S and a 4Fe-4S. - Escherichia coli anaerobic glycerol-3-phosphate dehydrogenase (EC 1.1.99.5) This enzyme is composed of three subunits: A, B, and C. The C subunit seems to be an iron-sulfur protein with two ferredoxin-like domains in the N-terminal part of the protein. - Escherichia coli anaerobic dimethyl sulfoxide reductase. The B subunit of this enzyme (gene dmsB) is an iron-sulfur protein with four 4Fe-4S ferredoxin-like domains. - Escherichia coli formate hydrogenlyase. Two of the subunits of this oligomeric complex (genes hycB and hycF) seem to be iron-sulfur proteins that each contain two 4Fe-4S ferredoxin-like domains. - Methanobacterium formicicum formate dehydrogenase (EC

1.2.1.2). This enzyme is used by the archaebacteria to grow on formate. The beta chain of this dimeric enzyme probably binds two 4Fe-4S centers. - *Escherichia coli* formate dehydrogenases N and O (EC 1.2.1.2). The beta chain of these two enzymes (genes *fdnH* and *fdoH*) are iron-sulfur proteins with four 4Fe-4S ferredoxin-like domains. - *Desulfovibrio* periplasmic [Fe] hydrogenase (EC 1.18.99.1). The large chain of this dimeric enzyme binds three 4Fe-4S centers, two of which are located in the ferredoxin-like N-terminal region of the protein. - *Methanobacterium thermoautrophicum* methyl viologen-reducing hydrogenase subunit *mvhB*, which contains six tandemly repeated ferredoxin-like domains and which probably binds twelve 4Fe-4S centers. - *Salmonella typhimurium* anaerobic sulfite reductase (EC 1.8.1.-) [4]. Two of the subunits of this enzyme (genes *asrA* and *asrC*) seem to both bind two 4Fe-4S centers. - A Ferredoxin-like protein (gene *fixX*) from the nitrogen-fixation genes locus of various *Rhizobium* species, and one from the *Nif*-region of *Azotobacter* species. - The 9 Kd polypeptide of chloroplast photosystem I [5] (gene *psaC*). This protein contains two low potential 4Fe-4S centers, referred as the A and B centers. - The chloroplast *frxB* protein which is predicted to carry two 4Fe-4S centers. - An ferredoxin from a primitive eukaryote, the enteric amoeba *Entamoeba histolytica*. - *Escherichia coli* hypothetical protein *yjjW*, a protein with a N-terminal region belonging to the radical activating enzymes family (see <PDOC00834>) and two potential 4Fe-4S centers. The pattern of cysteine residues in the iron-sulfur region is sufficient to detect this class of 4Fe-4S binding proteins.

Consensus pattern: C-x(2)-C-x(2)-C-x(3)-C-[PEG] [The four C's are 4Fe-4S ligands]-

[1] Meyer J. Trends Ecol. Evol. 3:222-226(1988).

[2] Otake E., Ooi T. J. Mol. Evol. 26:257-267(1987).

[3] Beinert H. FASEB J. 4:2483-2492(1990).

[4] Huang C.J., Barrett E.L. J. Bacteriol. 173:1544-1553(1991).

[5] Knaff D.B. Trends Biochem. Sci. 13:460-461(1988).

205. NifH/frxC family signatures (fer4_NifH)

Nitrogenase (EC 1.18.6.1) [1] is the enzyme system responsible for biological nitrogen fixation. Nitrogenase is an oligomeric complex which consists of two components: component 1 which contains the active site for the reduction of nitrogen to ammonia and

component 2 (also called the iron protein). Component 2 is a homodimer of a protein (gene *nifH*) which binds a single 4Fe-4S iron sulfur cluster [2]. In the nitrogen fixation process *nifH* is first reduced by a protein such as ferredoxin; the reduced protein then transfers electrons to component 1 with the concomitant consumption of ATP. A number of proteins are known to be evolutionary related to *nifH*. These proteins are: - Chloroplast encoded *frxC* (or *chlL*) protein [3]. *FrxC* is encoded on the chloroplast genome of some plant species, its exact function is not known, but it could act as an electron carrier in the conversion of protochlorophyllide to chlorophyllide. - *Rhodobacter capsulatus* proteins *bchL* and *bchX* [4]. These proteins are also likely to play a role in chlorophyll synthesis. There are a number of conserved regions in the sequence of these proteins: in the N-terminal section there is an ATP-binding site motif 'A' (P-loop) and in the central section there are two conserved cysteines which have been shown, in *nifH*, to be the ligands of the 4Fe-4S cluster. Two signature patterns that correspond to the regions around these cysteines were developed.

Consensus pattern: E-x-G-G-P-x(2)-[GA]-x-G-C-[AG]-G [C binds the iron-sulfur center]-
 Consensus pattern: D-x-L-G-D-V-V-C-G-G-F-[AG]-x-P [C binds the iron-sulfur center]-

[1] Pau R.N. Trends Biochem. Sci. 14:183-186(1989).

[2] Georgiadis M.M., Komiya H., Chakrabarti P., Woo D., Kornuc J.J., Rees D.C. Science 257:1653-1659(1992).

[3] Fujita Y., Takahashi Y., Kohchi T., Ozeki H., Ohyama K., Matsubara H. Plant Mol. Biol. 13:551-561(1989).

[4] Burke D.H., Alberti M., Hearst J.E. J. Bacteriol. 175:2407-2413(1993).

206. Ferritin iron-binding regions signatures

Ferritin [1,2] is one of the major non-heme iron storage proteins. It consists of a mineral core of hydrated ferric oxide, and a multi-subunit protein shell which englobes the former and assures its solubility in an aqueous environment. In animals the protein is mainly cytoplasmic and there are generally two or more genes that encode for closely related subunits (in mammals there are two subunits which are known as H(eavy) and L(ight)). In plants ferritin is found in the chloroplast [3]. There are a number of well conserved regions in the sequence of ferritins. Two of these regions to develop signature patterns were selected. The first pattern is

located in the central part of the sequence of ferritin and it contains three conserved glutamate which are thought to be involved in the binding of iron. The second pattern is located in the C-terminal section, it corresponds to a region which forms a hydrophilic channel through which small molecules and ions can gain access to the central cavity of the molecule; this pattern also includes conserved acidic residues which are potential metal-binding sites.

Consensus pattern: E-x-[KR]-E-x(2)-E-[KR]-[LF]-[LIVMA]-x(2)-Q-N-x-R-x-G-R [The 3 E's are potential iron ligands]-

Consensus pattern: D-x(2)-[LIVMF]-[STAC]-[DH]-F-[LI]-[EN]-x(2)-[FY]-L-x(6)-[LIVM]-[KN] [The second D and the E are potential iron ligands]-

[1] Crichton R.R., Charlotiaux-Wauters M. Eur. J. Biochem. 164:485-506(1987).

[2] Theil E.C. Annu. Rev. Biochem. 56:289-315(1987).

[3] Ragland M., Briat J.-F., Gagnon J., Laulhere J.-P., Massenet O., Theil E.C. J. Biol. Chem. 265:18339-18344(1990).

207. Intermediate filaments signature (filament)

Intermediate filaments (IF) [1,2,3] are proteins which are primordial components of the cytoskeleton and the nuclear envelope. They generally form filamentous structures 8 to 14 nm wide. IF proteins are members of a very large multigene family of proteins which has been subdivided in five major subgroups: - Type I: Acidic cytokeratins. - Type II: Basic cytokeratins. - Type III: Vimentin, desmin, glial fibrillary acidic protein (GFAP), peripherin, and plasticin. - Type IV: Neurofilaments L, H and M, alpha-internexin and nestin. - Type V: Nuclear lamins A, B1, B2 and C. All IF proteins are structurally similar in that they consist of: a central rod domain comprising some 300 to 350 residues which is arranged in coiled-coiled alpha-helices, with at least two short characteristic interruptions; a N-terminal non-helical domain (head) of variable length; and a C-terminal domain (tail) which is also non-helical, and which shows extreme length variation between different IF proteins. While IF proteins are evolutionary and structurally related, they have limited sequence homologies except in several regions of the rod domain. A conserved region at the C-terminal extremity of the rod domain was used as a sequence pattern for this class of proteins.

Consensus pattern: [IV]-x-[TACI]-Y-[RKH]-x-[LM]-L-[DE]-

[1] Quinlan R., Hutchison C., Lane B. Protein Prof. 2:801-952(1995).

[2] Steiner P.M., Roop D.R. Annu. Rev. Biochem. 57:593-625(1988).

5 [3] Stewart M. Curr. Opin. Cell Biol. 2:91-100(1990).

208. Flavodoxin signature

10 Flavodoxins [1,E1] are electron-transfer proteins that function in various electron transport systems. Flavodoxins bind one FMN molecule, which serves as a redox-active prosthetic group. Flavodoxins are functionally interchangeable with ferredoxins. They have been isolated from prokaryotes, cyanobacteria, and some eukaryotic algae. The signature pattern for these proteins is derived from a conserved region in their N-terminal section, this region is involved in the binding of the FMN phosphate group.

15 Consensus pattern: [LIV]-[LIVFY]-[FY]-x-[ST]-x(2)-[AGC]-x-T-x(3)-A-x(2)-[LIV]-

[1] Wakabayashi S., Kimura K., Matsubara H., Rogers L.J. Biochem. J. 263:981-984(1989).

209. Growth factor and cytokines receptors family signatures (fn3)

25 A number of receptors for lymphokines, hematopoietic growth factors and growth hormone-related molecules have been found [1 to 5] to share a common binding domain. Receptors known to belong to this family are: - Cytokine receptor common beta chain. This chain is common to the IL-3, IL-5 and GM-CSF receptors. - Cytokine receptor common gamma chain. This chain is common to the IL-2, IL-4, IL-7 and IL-13 receptors. - Ciliary neurotrophic factor receptor (CNTFR). - Erythropoietin receptor (EPOR). - Granulocyte colony-stimulating factor receptor (G-CSFR). - Granulocyte-macrophage colony-stimulating factor receptor alpha chain (GM-CSFR). - Interleukin-2 receptor beta chain (IL2R-beta). -
30 Interleukin-3 receptor alpha chain (IL3R). - Interleukin-4 receptor alpha chain (IL4R). - Interleukin-5 receptor alpha chain (IL5R). - Interleukin-6 receptor (IL6R). - Interleukin-7 receptor alpha chain (IL7R). - Interleukin-9 receptor (IL9R). - Growth hormone receptor (GRHR). - Prolactin receptor (PRLR). - Thrombopoietin receptor (TPOR). The conserved

region constitutes all or part of the extracellular ligand-binding region and is about 200 amino acid residues long. In the N-terminal of this domain there are two pairs of cysteines known, in the growth hormone receptor, to be involved in disulfide bonds. +-----

-----xxxxxxx-----+ | C C C C Extracellular XXXXXXXX Cytoplasmic | +-
5 |-----|-----xxxxxxx-----+ ||| Transmembrane +-+ +-+

+ Two patterns to detect this family of receptors were used. The first one is derived from the first N-terminal disulfide loop, the second is a tryptophan-rich pattern located at the C-terminal extremity of the extracellular region.

10 Consensus pattern: C-[LVFYR]-x(7,8)-[STIVDN]-C-x-W [The two C's are linked by a disulfide bond]-

Consensus pattern: [STGL]-x-W-[SG]-x-W-S-

[1] Bazan J.F. Biochem. Biophys. Res. Commun. 164:788-795(1989).

15 [2] Bazan J.F. Proc. Natl. Acad. Sci. U.S.A. 87:6934-6938(1990).

[3] Cosman D., Lyman S.D., Idzerda R.L., Beckmann M.P., Park L.S., Goodwin R.G., March C.J. Trends Biochem. Sci. 15:265-270(1990).

[4] d'Andrea A.D., Fasman G.D., Lodish H.F. Cell 58:1023-1024(1989).

[5] d'Andrea A.D., Fasman G.D., Lodish H.F. Curr. Opin. Cell Biol. 2:648-651(1990).

20 210. Phosphoribosylglycinamide formyltransferase active site (formyl_transf)

Phosphoribosylglycinamide formyltransferase (EC 2.1.2.2) (GART) [1] catalyzes the third step in de novo purine biosynthesis, the transfer of a formyl group to 5'-

25 phosphoribosylglycinamide. In higher eukaryotes, GART is part of a multifunctional enzyme polypeptide that catalyzes three of the steps of purine biosynthesis. In bacteria, plants and yeast, GART is a monofunctional protein of about 200 amino-acid residues. In the Escherichia coli enzyme, an aspartic acid residue has been shown to be involved in the catalytic mechanism. The region around this active site residue is well conserved in GART

30 from prokaryotic and eukaryotic sources and can be used as a signature pattern. Mammalian formyltetrahydrofolate dehydrogenase (EC 1.5.1.6) [2] is a cytosolic enzyme responsible for the NADP-dependent decarboxylative reduction of 10-formyltetrahydrofolate into tetrahydrofolate. It is a protein of about 900 amino acids consisting of three domains; the N-

terminal domain (200 residues) is structurally related to GARTs. Escherichia coli methionyl-tRNA formyltransferase (EC 2.1.2.9) (gene fmt) [3] is the enzyme responsible for modifying the free amino group of the aminoacyl moiety of methionyl-tRNA (fMet). The central part of fmt seems to be evolutionary related to GART's active site region.

5

Consensus pattern: G-x-[STM]-[IVT]-x-[FYWVQ]-[VMAT]-x-[DEVN]-x-[LIVMY]-D-x-G-x(2)-[LIVT]-x(6)-[LIVM] [D is the active site residue] -

[1] Inglese J., Smith J.M., Benkovic S.J. Biochemistry 29:6678-6687(1990).

10 [2] Cook R.J., Lloyd R.S., Wagner C. J. Biol. Chem. 266:4965-4973(1991).

[3] Guillon J.-M., Mechulam Y., Schmitter J.-M., Blanquet S., Fayat G. J. Bacteriol. 174:4294-4301(1992).

15

211. G10 protein signatures

A Xenopus protein known as G10 [1] has been found to be highly conserved in a wide range of eukaryotic species. The function of G10 is still unknown. G10 is a protein of about 17 to 18 Kd (143 to 157 residues) which is hydrophilic and whose C-terminal half is rich in cysteines and could be involved in metal-binding. As signature patterns, two of these cysteine-rich segments were selected.

20

Consensus pattern: L-C-C-x-[KR]-C-x(4)-[DE]-x-N-x(4)-C-x-C-R-V-P-

Consensus pattern: C-x-H-C-G-C-[KRH]-G-C-[SA]-

25 [1] McGrew L.L., Dworkin-Rastl E., Dworkin M.B., Richter J.D. Genes Dev. 3:803-815(1989).

212. G-protein alpha subunit

30

G proteins couple receptors of extracellular signals to intracellular signaling pathways. The G protein alpha subunit binds guanyl nucleotide and is a weak GTPase. Number of members: 195

[1] Coleman DE, Berghuis AM, Lee E, Linder ME, Gilman AG, Sprang SR, Science 1994;265:1405-1412.

[2] How G proteins work: a continuing story. Coleman DE, Sprang SR, Trends Biochem Sci 1996;21:41-44.

5

213. Glucose-6-phosphate dehydrogenase active site (G6PD)

Glucose-6-phosphate dehydrogenase (EC 1.1.1.49) (G6PD) [1] catalyzes the first step in the pentose pathway, the reduction of glucose-6-phosphate to gluconolactone 6-phosphate. A lysine residue has been identified as are active nucleophile associated with the activity of the enzyme. The sequence around this lysine is totally conserved from bacterial to mammalian G6PD's and can be used as a signature pattern

10

Consensus pattern: D-H-Y-L-G-K-[EQK] [K is the active site residue]-

15

[1] Jeffery J., Persson B., Wood I., Bergman T., Jeffery R., Joernvall H. Eur. J. Biochem. 212:41-49(1993).

20

214. GATA-type zinc finger domain

The GATA family of transcription factors are proteins that bind to DNA sites with the consensus sequence (A/T)GATA(A/G), found within the regulatory region of a number of genes. Proteins currently known to belong to this family are: - GATA-1 [1] (also known as Eryf1, GF-1 or NF-E1), which binds to the GATA region of globin genes and other genes expressed in erythroid cells. It is a transcriptional activator which probably serves as a general 'switch' factor for erythroid development. - GATA-2 [2], a transcriptional activator which regulates endothelin-1 gene expression in endothelial cells. - GATA-3 [3], a transcriptional activator which binds to the enhancer of the T-cell receptor alpha and delta genes. - GATA-4 [4], a transcriptional activator expressed in endodermally derived tissues and heart. - Drosophila protein pannier (or DGATAa) (gene pnr) which acts as a repressor of the achaete-scute complex (as-c). - Bombyx mori BCFI [5], which regulates the expression of chorion genes. - Caenorhabditis elegans elt-1 and elt-2, transcriptional activators of genes containing the GATA region, including vitellogenin genes [6]. - Ustilago maydis urbs1 [7], a

25

30

protein involved in the repression of the biosynthesis of siderophores. - Fission yeast protein GAF2. All these transcription factors contain a pair of highly similar 'zinc finger' type domains with the consensus sequence C-x₂-C-x₁₇-C-x₂-C. Some other proteins contain a single zinc finger motif highly related to those of the GATA transcription factors. These

5 proteins are: - *Drosophila* box A-binding factor (ABF) (also known as protein serpent (gene *srp*)) which may function as a transcriptional activator protein and may play a key role in the organogenesis of the fat body. - *Emericella nidulans* areA [8], a transcriptional activator which mediates nitrogen metabolite repression. - *Neurospora crassa* nit-2 [9], a transcriptional activator which turns on the expression of genes coding for enzymes required

10 for the use of a variety of secondary nitrogen sources, during conditions of nitrogen limitation. - *Neurospora crassa* white collar proteins 1 and 2 (WC-1 and WC-2), which control expression of light-regulated genes. - *Saccharomyces cerevisiae* DAL81 (or UGA43), a negative nitrogen regulatory protein. - *Saccharomyces cerevisiae* GLN3, a positive nitrogen regulatory protein. - *Saccharomyces cerevisiae* GAT1. - *Saccharomyces cerevisiae* GZF3.

Consensus pattern: C-x-[DN]-C-x(4,5)-[ST]-x(2)-W-[HR]-[RK]-x(3)-[GN]-x(3,4)- C-N-[AS]-C [The four C's are zinc ligands]

[1] Trainor C.D., Evans T., Felsenfeld G., Boguski M.S. *Nature* 343:92-96(1990).

[2] Lee M.E., Temizer D.T., Clifford J.A., Quertermous T. *J. Biol. Chem.* 266:16188-16192(1991).

[3] Ho I.-C., Vorhees P., Marin N., Oakley B.K., Tsai S.-F., Orkin S.H., Leiden J.M. *EMBO J.* 10:1187-1192(1991).

[4] Spieth J., Shim Y.H., Lea K., Conrad R., Blumenthal T. *Mol. Cell. Biol.* 11:4651-4659(1991).

[5] Drevet J.R., Skeiky Y.A., Iatrou K. *J. Biol. Chem.* 269:10660-10667(1994).

[6] Hawkins M.G., McGhee J.D. *J. Biol. Chem.* 270:14666-14671(1995).

[7] Voisard C.P.O., Wang J., Xu P., Leong S.A., McEvoy J.L. *Mol. Cell. Biol.* 13:7091-7100(1993).

[8] Arst H.N. Jr., Kudla B., Martinez-Rossi N.M., Caddick M.X., Sibley S., Davies R.W. *Trends Genet.* 5:291-291(1989).

[9] Fu Y.-H., Marzluf G.A. *Mol. Cell. Biol.* 10:1056-1065(1990).

215. Glutamine amidotransferases class-I active site (GATase)

A large group of biosynthetic enzymes are able to catalyze the removal of the ammonia group from glutamine and then to transfer this group to a substrate to form a new carbon-nitrogen group. This catalytic activity is known as glutamine amidotransferase (GATase) (EC 2.4.2.-) [1]. The GATase domain exists either as a separate polypeptidic subunit or as part of a larger polypeptide fused in different ways to a synthase domain. On the basis of sequence similarities two classes of GATase domains have been identified [2,3]: class-I (also known as trpG-type) and class-II (also known as purF-type). Class-I GATase domains have been found in the following enzymes: - The second component of anthranilate synthase (AS) (EC 4.1.3.27) [4]. AS catalyzes the biosynthesis of anthranilate from chorismate and glutamine. AS is generally a dimeric enzyme: the first component can synthesize anthranilate using ammonia rather than glutamine, whereas component II provides the GATase activity. In some bacteria and in fungi the GATase component of AS is part of a multifunctional protein that also catalyzes other steps of the biosynthesis of tryptophan. - The second component of 4-amino-4-deoxychorismate (ADC) synthase (EC 4.1.3. -), a dimeric prokaryotic enzyme that function in the pathway that catalyzes the biosynthesis of para-aminobenzoate (PABA) from chorismate and glutamine. The second component (gene pabA) provides the GATase activity [4]. - CTP synthase (EC 6.3.4.2). CTP synthase catalyzes the final reaction in the biosynthesis of pyrimidine, the ATP-dependent formation of CTP from UTP and glutamine. CTP synthase is a single chain enzyme that contains two distinct domains; the GATase domain is in the C-terminal section [2]. - GMP synthase (glutamine-hydrolyzing) (EC 6.3.5.2). GMP synthase catalyzes the ATP-dependent formation of GMP from xanthosine 5'-phosphate and glutamine. GMP synthase is a single chain enzyme that contains two distinct domains; the GATase domain is in the N-terminal section [5]. - Glutamine-dependent carbamoyl-phosphate synthase (EC 6.3.5.5) (GD-CPSase); an enzyme involved in both arginine and pyrimidine biosynthesis and which catalyzes the ATP-dependent formation of carbamoyl phosphate from glutamine and carbon dioxide. In bacteria GD-CPSase is composed of two subunits: the large chain (gene carB) provides the CPSase activity, while the small chain (gene carA) provides the GATase activity. In yeast the enzyme involved in arginine biosynthesis is also composed of two subunits: CPA1 (GATase), and CPA2 (CPSase). In most eukaryotes, the first three steps of pyrimidine biosynthesis are catalyzed by a large multifunctional enzyme (called URA2 in yeast, rudimentary in Drosophila, and CAD

in mammals). The GATase domain is located at the N-terminal extremity of this polypeptide [6]. - Phosphoribosylformylglycinamidine synthase II (EC 6.3.5.3), an enzyme that catalyzes the fourth step in the de novo biosynthesis of purines. In some species of bacteria, FGAM synthase II is composed of two subunits: a small chain (gene purQ) which provides the GATase activity and a large chain (gene purL) which provides the aminator activity. - The histidine amidotransferase hisH, an enzyme that catalyzes the fifth step in the biosynthesis of histidine in prokaryotes. In the second component of AS a cysteine has been shown [7] to be essential for the amidotransferase activity. The sequence around this residue is well conserved in all the above GATase domains and can be used as a signature pattern for class-I GATase.-

Consensus pattern: [PAS]-[LIVMFYT]-[LIVMFY]-G-[LIVMFY]-C-[LIVMFYN]-G-x-[QEH]-x-[LIVMFA] [C is the active site residue]-

[1] Buchanan J.M. Adv. Enzymol. 39:91-183(1973).

[2] Weng M., Zalkin H. J. Bacteriol. 169:3023-3028(1987).

[3] Nyunoya H., Lusty C.J. J. Biol. Chem. 259:9790-9798(1984).

[4] Crawford I.P. Annu. Rev. Microbiol. 43:567-600(1989).

[5] Zalkin H., Argos P., Narayana S.V.L., Tiedeman A.A., Smith J.M. J. Biol. Chem. 260:3350-3354(1985).

[6] Davidson J.N., Chen K.C., Jamison R.S., Musmanno L.A., Kern C.B. BioEssays 15:157-164(1993).

[7] Tso J.Y., Hermodson M.A., Zalkin H. J. Biol. Chem. 255:1451-1457(1980).

216. Glutamine amidotransferases class-II active site (GATase_2)

A large group of biosynthetic enzymes are able to catalyze the removal of the ammonia group from glutamine and then to transfer this group to a substrate to form a new carbon-nitrogen group. This catalytic activity is known as glutamine amidotransferase (GATase) (EC 2.4.2.-) [1]. The GATase domain exists either as a separate polypeptidic subunit or as part of a larger polypeptide fused in different ways to a synthase domain. On the basis of sequence similarities two classes of GATase domains have been identified [2,3]: class-I(also known as trpG-type) and class-II (also known as purF-type). Class-II GATase domains have been found in the following enzymes: - Amido phosphoribosyltransferase (glutamine

phosphoribosylpyrophosphate amidotransferase) (EC 2.4.2.14). An enzyme which catalyzes the first step in purine biosynthesis, the transfer of the ammonia group of glutamine to PRPP to form 5-phosphoribosylamine (gene purF in bacteria, ADE4 in yeast). - Glucosamine--fructose-6-phosphate aminotransferase (EC 2.6.1.16). This enzyme catalyzes a key reaction in amino sugar synthesis, the formation of glucosamine 6-phosphate from fructose 6-phosphate and glutamine (gene glmS in Escherichia coli, nodM in Rhizobium, GFA1 in yeast) - Asparagine synthetase (glutamine-hydrolyzing) (EC 6.3.5.4). This enzyme is responsible for the synthesis of asparagine from aspartate and glutamine. A cysteine is present at the N-terminal extremity of the mature form of all these enzymes. The cysteine has been shown, in amido phosphoribosyltransferase [4] and in asparagine synthetase [5] to be important for the catalytic mechanism.

Consensus pattern: <x(0,11)-C-[GS]-[IV]-[LIVMFYW]-[AG] [C is the active site residue]-

- [1] Buchanan J.M. Adv. Enzymol. 39:91-183(1973).
- [2] Weng M., Zalkin H. J. Bacteriol. 169:3023-3028(1987).
- [3] Nyunoya H., Lusty C.J. J. Biol. Chem. 259:9790-9798(1984).
- [4] van Heeke G., Schuster M. J. Biol. Chem. 264:5503-5509(1989).
- [5] Vollmer S.J., Switzer R.L., Hermodson M.A., Bower S.G., Zalkin H. J. Biol. Chem. 258:10582-10585(1983).

217. GDP dissociation inhibitor (GDI)

- [1] Schalk I, Zeng K, Wu SK, Stura EA, Matteson J, Huang M, Tandon A, Wilson IA, Balch WE, Nature 1996;381:42-48.

218. Oxidoreductase family (GFO_IDH_MocA)

This family of enzymes utilise NADP or NAD. This family: is called the GFO/IDH/MOCA family in swiss-prot.

- [1] Kingston RL, Scopes RK, Baker EN, Structure 1996;4:1413-1428.

245

219. GHMP kinases putative ATP-binding domain

The following kinases contains, in their N-terminal section, a conserved Gly/Ser-rich region which is probably involved in the binding of ATP [1]. These kinases are listed below. -

Galactokinase (EC 2.7.1.6). - Homoserine kinase (EC 2.7.1.39). - Mevalonate kinase (EC 2.7.1.36). - Phosphomevalonate kinase (EC 2.7.4.2). This group of kinases was called 'GHMP' (from the first letter of their substrate)

Consensus pattern: [LIVM]-[PK]-x-[GSTA]-x(0,1)-G-L-[GS]-S-S-[GSA]-[GSTAC]-

[1] Tsay Y.H., Robinson G.W. Mol. Cell. Biol. 11:620-631(1991).

220. Glucose inhibited division protein A family signatures (GIDA)

Bacterial glucose inhibited division protein A (gene gidA) is a protein of 70Kd whose function is not yet known and whose sequence is highly conserved. It is evolutionary related to yeast hypothetical protein YGL236C, Caenorhabditis elegans hypothetical protein F52H3.2 and a Bacillus subtilis protein called gid (and which is different from B.subtilis gidA). Two highly conserved regions were selected as signature patterns. Both regions are located in the central region of the protein.

Consensus pattern: [GS]-[PT]-x-Y-C-P-S-[LIVM]-E-x-K-[LIVM]-x-[KR]-

Consensus pattern: A-G-Q-x-[NT]-G-x(2)-G-Y-x-E-[SAG](3)-[QS]-G-[LIVM](2)-A-G-[LIVMT]-N-A-

221. (GLFV_dehydrog)

Glu / Leu / Phe / Val dehydrogenases active site

- Glutamate dehydrogenases (EC 1.4.1.2, EC 1.4.1.3, and EC 1.4.1.4) (GluDH)

are enzymes that catalyze the NAD- or NADP-dependent reversible deamination of glutamate into alpha-ketoglutarate [1,2]. GluDH isozymes are generally involved with either ammonia assimilation or glutamate catabolism.

- Leucine dehydrogenase (EC 1.4.1.9) (LeuDh) is a NAD-dependent enzyme that

catalyzes the reversible deamination of leucine and several other aliphatic amino acids to their keto analogues [3].

- Phenylalanine dehydrogenase (EC 1.4.1.20) (PheDH) is a NAD-dependent enzyme that catalyzes the reversible deamidation of L-phenylalanine into phenyl-pyruvate [4].
- Valine dehydrogenase (EC 1.4.1.8) (ValDH) is a NADP-dependent enzyme that catalyzes the reversible deamidation of L-valine into 3-methyl-2-oxobutanoate [5].

These dehydrogenases are structurally and functionally related. A conserved lysine residue located in a glycine-rich region has been implicated in the catalytic mechanism. The conservation of the region around this residue allows the derivation of a signature pattern for such type of enzymes.

Consensus pattern[LIV]-x(2)-G-G-[SAG]-K-x-[GV]-x(3)-[DNST]-[PL] [K is the active site residue] Sequences known to belong to this class detected by the pattern ALL.

Note all known sequences from this family have Pro in the last position of the pattern with the exception of yeast GluDH which as Leu.

[1] Britton K.L., Baker P.J., Rice D.W., Stillman T.J. Eur. J. Biochem. 209:851-859(1992).

[2] Benachenhou-Lahfa N., Forterre P., Labedan B. J. Mol. Evol. 36:335-346(1993).

[3] Nagata S., Tanizawa K., Esaki N., Sakamoto Y., Ohshima T., Tanaka H., Soda K. Biochemistry 27:9056-9062(1988).

[4] Takada H., Yoshimura T., Ohshima T., Esaki N., Soda K. J. Biochem. 109:371-376(1991).

[5] Hutchinson C.R., Tang L. J. Bacteriol. 175:4176-4185(1993).

222. GMC oxidoreductases signatures

The following FAD flavoproteins oxidoreductases have been found [1,2] to be evolutionary related. These enzymes, which are called 'GMC oxidoreductases', are listed below. - Glucose oxidase (EC 1.1.3.4) (GOX) from *Aspergillus niger*. Reaction catalyzed: glucose + oxygen ->

delta-gluconolactone + hydrogen peroxide. - Methanol oxidase (EC 1.1.3.13) (MOX) from fungi. Reaction catalyzed: methanol + oxygen -> acetaldehyde + hydrogen peroxide. - Choline dehydrogenase (EC 1.1.99.1) (CHD) from bacteria. Reaction catalyzed: choline + unknown acceptor -> betaine acetaldehyde + reduced acceptor. - Glucose dehydrogenase (GLD) (EC 1.1.99.10) from Drosophila. Reaction catalyzed: glucose + unknown acceptor -> delta-gluconolactone + reduced acceptor. - Cholesterol oxidase (CHOD) (EC 1.1.3.6) from Brevibacterium sterolicum and Streptomyces strain SA-COO. Reaction catalyzed: cholesterol + oxygen -> cholest-4-en-3-one + hydrogen peroxide. - AlkJ [3], an alcohol dehydrogenase from Pseudomonas oleovorans, which converts aliphatic medium-chain-length alcohols into aldehydes. This family also includes a lyase: - (R)-mandelonitrile lyase (EC 4.1.2.10) (hydroxynitrile lyase) from plants [4], an enzyme involved in cyanogenesis, the release of hydrogen cyanide from injured tissues. These enzymes are proteins of size ranging from 556 (CHD) to 664 (MOX) amino acid residues which share a number of regions of sequence similarities. One of these regions, located in the N-terminal section, corresponds to the FAD ADP-binding domain. The function of the other conserved domains is not yet known; two of these domains were selected as signature patterns. The first one is located in the N-terminal section of these enzymes, about 50 residues after the ADP-binding domain, while the second one is located in the central section.

Consensus pattern: [GA]-[RKN]-x-[LIV]-G(2)-[GST](2)-x-[LIVM]-N-x(3)-[FYWA]-x(2)-[PAG]-x(5)-[DNESH]-

Consensus pattern: [GS]-[PSTA]-x(2)-[ST]-P-x-[LIVM](2)-x(2)-S-G-[LIVM]-G-

[1] Cavener D.R. J. Mol. Biol. 223:811-814(1992).

[2] Henikoff S., Henikoff J.G. Genomics 19:97-107(1994).

[3] van Beilen J.B., Eggink G., Enequist H., Bos R., Witholt B. Mol. Microbiol. 6:3121-3136(1992).

[4] Cheng I.P., Poulton J.E. Plant Cell Physiol. 34:1139-1143(1993).

223. (GMP_synt_C)

Glutamine amidotransferases class-I active site

A large group of biosynthetic enzymes are able to catalyze the removal of the ammonia group from glutamine and then to transfer this group to a substrate to form a new carbon-nitrogen group. This catalytic activity is known as glutamine amidotransferase (GATase) (EC 2.4.2.-) [1]. The GATase domain exists either as a separate polypeptidic subunit or as part of a larger polypeptide fused in different ways to a synthase domain. On the basis of sequence similarities two classes of GATase domains have been identified [2,3]: class-I (also known as trpG-type) and class-II (also known as purF-type). Class-I GATase domains have been found in the following enzymes:

- The second component of anthranilate synthase (AS) (EC 4.1.3.27) [4]. AS catalyzes the biosynthesis of anthranilate from chorismate and glutamine. AS is generally a dimeric enzyme: the first component can synthesize anthranilate using ammonia rather than glutamine, whereas component II provides the GATase activity. In some bacteria and in fungi the GATase component of AS is part of a multifunctional protein that also catalyzes other steps of the biosynthesis of tryptophan.

- The second component of 4-amino-4-deoxychorismate (ADC) synthase (EC 4.1.3. -), a dimeric prokaryotic enzyme that function in the pathway that catalyzes the biosynthesis of para-aminobenzoate (PABA) from chorismate and glutamine. The second component (gene pabA) provides the GATase activity [4].

- CTP synthase (EC 6.3.4.2). CTP synthase catalyzes the final reaction in the biosynthesis of pyrimidine, the ATP-dependent formation of CTP from UTP and glutamine. CTP synthase is a single chain enzyme that contains two distinct domains; the GATase domain is in the C-terminal section [2].

- GMP synthase (glutamine-hydrolyzing) (EC 6.3.5.2). GMP synthase catalyzes the ATP-dependent formation of GMP from xanthosine 5'-phosphate and glutamine. GMP synthase is a single chain enzyme that contains two distinct domains; the GATase domain is in the N-terminal section [5].

- Glutamine-dependent carbamoyl-phosphate synthase (EC 6.3.5.5) (GD-CPSase); an enzyme involved in both arginine and pyrimidine biosynthesis and which catalyzes the ATP-dependent formation of carbamoyl phosphate from glutamine and carbon dioxide. In bacteria GD-CPSase is composed of two subunits: the large chain (gene carB) provides the CPSase activity, while the small chain (gene carA) provides the GATase activity. In yeast the enzyme involved in arginine biosynthesis is also composed of two subunits: CPA1 (GATase),

and CPA2 (CPSase). In most eukaryotes, the first three steps of pyrimidine biosynthesis are catalyzed by a large multifunctional enzyme (called URA2 in yeast, rudimentary in *Drosophila*, and CAD in mammals). The GATase domain is located at the N-terminal extremity of this polypeptide [6].

5 - Phosphoribosylformylglycinamide synthase II (EC 6.3.5.3), an enzyme that catalyzes the fourth step in the de novo biosynthesis of purines. In some species of bacteria, FGAM synthase II is composed of two subunits: a small chain (gene *purQ*) which provides the GATase activity and a large chain (gene *purL*) which provides the aminator activity.

10 - The histidine amidotransferase *hisH*, an enzyme that catalyzes the fifth step in the biosynthesis of histidine in prokaryotes.

In the second component of AS a cysteine has been shown [7] to be essential for the amidotransferase activity. The sequence around this residue is well conserved in all the above GATase domains and can be used as a signature pattern for class-I GATase.

15 Consensus pattern[PAS]-[LIVMFYT]-[LIVMFY]-G-[LIVMFY]-C-[LIVMFYN]-G-x-[QEH]-x-[LIVMFA] [C is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for 6 sequences.

20 Note: in the first position of the pattern Pro is found in all cases except in the slime mold GD-CPSase where it is replaced by Ala.

[1] Buchanan J.M. Adv. Enzymol. 39:91-183(1973).

[2] Weng M., Zalkin H. J. Bacteriol. 169:3023-3028(1987).

25 [3] Nyunoya H., Lusty C.J. J. Biol. Chem. 259:9790-9798(1984).

[4] Crawford I.P. Annu. Rev. Microbiol. 43:567-600(1989).

[5] Zalkin H., Argos P., Narayana S.V.L., Tiedeman A.A., Smith J.M. J. Biol. Chem. 260:3350-3354(1985).

30 [6] Davidson J.N., Chen K.C., Jamison R.S., Musmanno L.A., Kern C.B. BioEssays 15:157-164(1993).

[7] Tso J.Y., Hermodson M.A., Zalkin H. J. Biol. Chem. 255:1451-1457(1980).

224. Glutathione peroxidases signatures (GSHPx)

Glutathione peroxidase (EC 1.11.1.9) (GSHPx) [1,2] is an enzyme that catalyzes the reduction of hydroxyperoxides by glutathione. Its main function is to protect against the damaging effect of endogenously formed hydroxyperoxides. In higher vertebrates at least four forms of GSHPx are known to exist: a ubiquitous cytosolic form (GSHPx-1), a gastrointestinal cytosolic for (GSHPx-GI) [3], a plasma secreted form (GSHPx-P) [4], and a epididymal secretory form (GSHPx-EP). In addition to these characterized forms, the sequence of a protein of unknown function [5] has been shown to be evolutionary related to those of GSHPx's. In filarial nematode parasites such as *Brugia pahangi* the major soluble cuticular protein, known as gp29, is a secreted GSHPx which could provide a mechanism of resistance to the immune reaction of the mammalian host by neutralizing the products of the oxidative burst of leukocytes [6]. *Escherichia coli* protein btuE, a periplasmic protein involved in the transport of vitamin B12, is also evolutionary related to GSHPx's; the significance of this relationship is not yet clear. Selenium, in the form of selenocysteine [7] is part of the catalytic site of GSHPx. The sequence around the selenocysteine residue is moderately well conserved in GSHPx's and the related proteins and can be used as a signature pattern. As a second signature for this family of proteins a highly conserved octapeptide located in the central section of these proteins was selected.

Consensus pattern: [GN]-[RKHNFYC]-x-[LIVMFC]-[LIVMF](2)-x-N-[VT]-x-[STC]-x-C-[GA]-x-T [C is the active site selenocysteine residue]

Consensus pattern: [LIV]-[AGD]-F-P-[CS]-[NG]-Q-

[1] Mannervik B. *Meth. Enzymol.* 113:490-495(1985).

[2] Mullenbach G.T., Tabrizi A., Irvine B.D., Bell G.I., Tainer J.A., Hallewell R.A. *Protein Eng.* 2:239-246(1988).

[3] Chu F.F., Doroshov J.H., Esworthy R.S. *J. Biol. Chem.* 268:2571-2576(1993).

[4] Takahashi K., Akasaka M., Yamamoto Y., Kobayashi C., Mizoguchi J., Koyama J. *Biochem.* 108:145-148(1990).

[5] Dunn D.K., Howells D.D., Richardson J., Goldfarb P.S. *Nucleic Acids Res.* 17:6390-6390(1989).

[6] Cookson E., Blaxter M.L., Selkirk M.E. *Proc. Natl. Acad. Sci. U.S.A.* 89:5837-5841(1992).

[7] Stadtman T.C. Annu. Rev. Biochem. 59:111-127(1990).

225. (GST)

5 Glutathione S-transferases

Function: conjugation of reduced glutathione to a variety of targets. Also included in the alignment, but are not GSTs S-crystallins from squid. Similarity to GST was previously noted. Eukaryotic elongation factors 1-gamma. Not known to have GST activity; similarity not previously recognized. Supported by HMM and manual alignment inspection. HSP26 family of stress-related proteins. including auxin-regulated proteins in plants and stringent starvation proteins in E. coli. Not known to have GST activity. Similarity not previously recognized. Supported by HMM and manual alignment inspection. Alignment spans entire protein.

226. GTP1/OBG family signature

A widespread family of GTP-binding proteins has been recently characterized [1,2]. This family currently includes: - Mouse and Xenopus protein DRG. - Human protein DRG2. - Drosophila protein 128up. - Fission yeast protein gtp1. - A Halobacterium cutirubrum hypothetical protein in a ribosomal protein gene cluster. - Bacillus subtilis protein obg. Obg has been experimentally shown to bind GTP. - Escherichia coli hypothetical protein yhbZ. - Haemophilus influenzae hypothetical protein HI0877. - Mycoplasma genitalium hypothetical protein MG384. - Yeast hypothetical protein YAL036c (FUN11). - Yeast hypothetical protein YGR173w. - Caenorhabditis elegans hypothetical protein C02F5.3. The function of the proteins that belong to this family is not yet known. They are polypeptides of about 40 to 48 Kd which contain the five small sequence elements characteristic of GTP-binding proteins [3]. As a signature pattern the region that correspond to the ATP/GTP B motif (also called G-3 in GTP-binding proteins) was selected.

Consensus pattern: D-[LIVM]-P-G-[LIVM](2)-[DEY]-[GN]-A-x(2)-G-x-G -

- [1] Sazuka T., Tomooka Y., Ikawa Y., Noda M., Kumar S. Biochem. Biophys. Res. Commun. 189:363-370(1992).
- [2] Hudson J.D., Young P.G. Gene 125:191-193(1993).
- [3] Bourne H.R., Sanders D.A., McCormick F. Nature 349:117-127(1991).

5

227. (GTP_EFTU1)

ATP/GTP-binding site motif A (P-loop)

From sequence comparisons and crystallographic data analysis it has been shown

10 [1,2,3,4,5,6] that an appreciable proportion of proteins that bind ATP or GTP share a number of more or less conserved sequence motifs. The best conserved of these motifs is a glycine-rich region, which typically forms a flexible loop between a beta-strand and an alpha-helix. This loop interacts with one of the phosphate groups of the nucleotide. This sequence motif is generally referred to as the 'A' consensus sequence [1] or the 'P-loop' [5]. There are numerous

15 ATP- or GTP-binding proteins in which the P-loop is found. Listed below are a number of protein families for which the relevance of the presence of such motif has been noted: - ATP synthase alpha and beta subunits (see <PDOC00137>). - Myosin heavy chains. - Kinesin heavy chains and kinesin-like proteins (see <PDOC00343>). - Dynamins and dynamin-like proteins (see <PDOC00362>). - Guanylate kinase (see <PDOC00670>). - Thymidine kinase (see <PDOC00524>). - Thymidylate kinase (see <PDOC01034>). - Shikimate kinase (see <PDOC00868>). - Nitrogenase iron protein family (nifH/frxC) (see <PDOC00580>). - ATP-binding proteins involved in 'active transport' (ABC transporters) [7] (see <PDOC00185>). - DNA and RNA helicases [8,9,10]. - GTP-binding elongation factors (EF-Tu, EF-1alpha, EF-G, EF-2, etc.). - Ras family of GTP-binding proteins (Ras, Rho, Rab, Ral, Ypt1, SEC4, etc.).

25 - Nuclear protein ran (see <PDOC00859>). - ADP-ribosylation factors family (see <PDOC00781>). - Bacterial dnaA protein (see <PDOC00771>). - Bacterial recA protein (see <PDOC00131>). - Bacterial recF protein (see <PDOC00539>). - Guanine nucleotide-binding proteins alpha subunits (Gi, Gs, Gt, G0, etc.). - DNA mismatch repair proteins mutS family (See <PDOC00388>). - Bacterial type II secretion system protein E (see <PDOC00567>). Not

30 all ATP- or GTP-binding proteins are picked-up by this motif. A number of proteins escape detection because the structure of their ATP-binding site is completely different from that of the P-loop. Examples of such proteins are the E1-E2 ATPases or the glycolytic kinases. In other ATP- or GTP-binding proteins the flexible loop exists in a slightly different form; this

is the case for tubulins or protein kinases. A special mention must be reserved for adenylate kinase, in which there is a single deviation from the P-loop pattern: in the last position Gly is found instead of Ser or Thr.

5 -Consensus pattern: [AG]-x(4)-G-K-[ST]-

[1] Walker J.E., Saraste M., Runswick M.J., Gay N.J. EMBO J. 1:945-951(1982).

[2] Moller W., Amons R. FEBS Lett. 186:1-7(1985).

[3] Fry D.C., Kuby S.A., Mildvan A.S. Proc. Natl. Acad. Sci. U.S.A. 83:907-911(1986).

10 [4] Dever T.E., Glynias M.J., Merrick W.C. Proc. Natl. Acad. Sci. U.S.A. 84:1814-1818(1987).

[5] Saraste M., Sibbald P.R., Wittinghofer A. Trends Biochem. Sci. 15:430-434(1990).

[6] Koonin E.V. J. Mol. Biol. 229:1165-1174(1993).

[7] Higgins C.F., Hyde S.C., Mimmack M.M., Gileadi U., Gill D.R., Gallagher M.P. J.

15 Bioenerg. Biomembr. 22:571-592(1990).

[8] Hodgman T.C. Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).

[9] Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K., Schnier J., Slonimski P.P. Nature 337:121-122(1989).

[10] Gorbalenya A.E., Koonin E.V., Donchenko A.P., Blinov V.M. Nucleic Acids Res. 20 17:4713-4730(1989).

GTP-binding elongation factors signature (GTP_EFTU2)

Elongation factors [1,2] are proteins catalyzing the elongation of peptide chains in protein biosynthesis. In both prokaryotes and eukaryotes, there are three distinct types of elongation factors, as described in the following table: -----

----- Eukaryotes Prokaryotes Function -----
 ----- EF-1alpha EF-Tu Binds GTP and an aminoacyl-tRNA; delivers the latter to the A site of ribosomes. EF-1beta EF-Ts Interacts with EF-1a/EF-Tu to displace GDP and thus allows the regeneration of GTP-EF-1a. EF-2 EF-G Binds GTP and peptidyl-tRNA and translocates the latter from the A site to the P site. -----

30 -----The GTP-binding elongation factor family also includes the following proteins: - Eukaryotic peptide chain release factor GTP-binding subunits [3]. These proteins interact with release factors that bind to ribosomes that have encountered a stop codon at their

decoding site and help them to induce release of the nascent polypeptide. The yeast protein was known as SUP2 (and also as SUP35, SUF12 or GST1) and the human homolog as GST1-Hs. - Prokaryotic peptide chain release factor 3 (RF-3) (gene prfC). RF-3 is a class-II RF, a GTP-binding protein that interacts with class I RFs (see <PDOC00607>) and enhance their activity [4]. - Prokaryotic GTP-binding protein lepA and its homolog in yeast (gene GUF1) and in *Caenorhabditis elegans* (ZK1236.1). - Yeast HBS1 [5]. - Rat statin S1 [6], a protein of unknown function which is highly similar to EF-1alpha. - Prokaryotic selenocysteine-specific elongation factor selB [7], which seems to replace EF-Tu for the insertion of selenocysteine directed by the UGA codon. - The tetracycline resistance proteins tetM/tetO [8,9] from various bacteria such as *Campylobacter jejuni*, *Enterococcus faecalis*, *Streptococcus mutans* and *Ureaplasma urealyticum*. Tetracycline binds to the prokaryotic ribosomal 30S subunit and inhibits binding of aminoacyl-tRNAs. These proteins abolish the inhibitory effect of tetracycline on protein synthesis. - *Rhizobium* nodulation protein nodQ [10]. - *Escherichia coli* hypothetical protein yihK [11]. In EF-1-alpha, a specific region has been shown [12] to be involved in a conformational change mediated by the hydrolysis of GTP to GDP. This region is conserved in both EF-1alpha/EF-Tu as well as EF-2/EF-G and thus seems typical for GTP-dependent proteins which bind non-initiator tRNAs to the ribosome. The pattern developed for this family of proteins include that conserved region.

Consensus pattern: D-[KRSTGANQFYW]-x(3)-E-[KRAQ]-x-[RKQD]-[GC]-[IVMK]-[ST]-[IV]-x(2)-[GSTACKRNQ]-

[1] Concise Encyclopedia Biochemistry, Second Edition, Walter de Gruyter, Berlin New-York (1988).

[2] Moldave K. Annu. Rev. Biochem. 54:1109-1149(1985).

[3] Stansfield I., Jones K.M., Kushnirov V.V., Dagkesamanskaya A.R., Poznyakovski A.I., Paushkin S.V., Nierras C.R., Cox B.S., Ter-Avanesyan M.D., Tuite M.F. EMBO J. 14:4365-4373(1995).

[4] Grentzmann G., Brechemier-Baey D., Heurgue-Hamard V., Buckingham R.H. J. Biol. Chem. 270:10595-10600(1995).

[5] Nelson R.J., Ziegelhoffer T., Nicolet C., Werner-Washburne M., Craig E.A. Cell 71:97-105(1992).

[6] Ann D.K., Moutsatsos I.K., Nakamura T., Lin H.H., Mao P.-L., Lee M.-J., Chin S., Liem R.K.H., Wang E. J. Biol. Chem. 266:10429-10437(1991).

[7] Forchhammer K., Leinfelder W., Bock A. Nature 342:453-456(1989).

[8] Manavathu E.K., Hiratsuka K., Taylor D.E. Gene 62:17-26(1988).

5 [9] Leblanc D.J., Lee L.N., Titmas B.M., Smith C.J., Tenover F.C. J. Bacteriol. 170:3618-3626(1988).

[10] Cervantes E., Sharma S.B., Maillet F., Vasse J., Truchet G., Rosenberg C. Mol. Microbiol. 3:745-755(1989).

10 [11] Plunkett G. III, Burland V.D., Daniels D.L., Blattner F.R. Nucleic Acids Res. 21:3391-3398(1993).

[12] Moller W., Schipper A., Amons R. Biochimie 69:983-989(1987).

228. GTP cyclohydrolase II.

15 GTP cyclohydrolase II catalyses the first committed step in the biosynthesis of riboflavin.

[1] Richter G, Ritz H, Katzenmeier G, Volk R, Kohnle A, Lottspeich F, Allendorf D, Bacher A, J Bacteriol 1993;175:4045-4051.

229. Galactose-1-phosphate uridyl transferase signatures (GalP_UDP_transf)

Galactose-1-phosphate uridyl transferase (EC 2.7.7.10) (galT) catalyzes the transfer of an uridyldiphosphate group on galactose (or glucose) 1-phosphate. During the reaction, the uridyl moiety links to a histidine residue. In the Escherichia coli enzyme, it has been shown
25 [1] that two histidine residues separated by a single proline residue are essential for enzyme activity. On the basis of sequence similarities, two apparently unrelated families seem to exist. Class-I enzymes are found in eukaryotes as well as some bacteria such as Escherichia coli or Streptomyces lividans, while class-II enzymes have been found so far only in bacteria such as Bacillus subtilis or Lactobacillus helveticus [2]. Signature patterns for both families
30 were developed. For class-I enzymes the signature is based on the active site residues. For class-II enzymes a region which also includes two conserved histidines was chosen.

Consensus pattern: F-E-N-[RK]-G-x(3)-G-x(4)-H-P-H-x-Q [The two H's are the active site residues]-

Consensus pattern: D-L-P-I-V-G-G-[ST]-[LIVM](2)-[SA]-H-[DEN]-H-[FY]-Q-G-G -

Note: class-I enzymes are structurally related to the HIT family of proteins (see

5 <[PDOC00694](#)

[1] Reichardt J.K.V., Berg P. Nucleic Acids Res. 16:9017-9026(1988).

[2] Mollet B., Pilloud N. J. Bacteriol. 173:4464-4473(1991).

10

230. Gamma-thionins family signature

The following small plant proteins are evolutionary related:

- Gamma-thionins from wheat endosperm (gamma-purothionins) and barley (gamma- hordothionins) which are toxic to animal cells and inhibit protein synthesis in cell free systems [1].
- A flower-specific thionin (FST) from tobacco [2].
- Antifungal proteins (AFP) from the seeds of Brassicaceae species such as radish, mustard, turnip and Arabidopsis thaliana [3].
- Inhibitors of insect alpha-amylases from sorghum [4].
- Probable protease inhibitor P322 from potato.
- A germination-related protein from cowpea [5].
- Anther-specific protein SF18 from sunflower [6]. SF18 is a protein that contains a gamma-thionin domain at its N-terminus and a proline-rich C- terminal domain.
- Soybean sulfur-rich protein SE60 [7].
- Vicia faba antibacterial peptides fabatin-1 and -2.

In their mature form, these proteins generally consist of about 45 to 50 amino-acid residues. As shown in the following schematic representation, these peptides contain eight conserved cysteines involved in disulfide bonds.

+-----+ | +-----+ |||||
 30 xxCxxxxxxxxxCxxxxCxxxCxxxxxxxxxCxxxxxCxCxxxC *****|***|||
 +---|-----+ | +-----+

'C': conserved cysteine involved in a disulfide bond.

'*': position of the pattern.

Consensus pattern: [KRG]-x-C-x(3)-[SV]-x(2)-[FYWH]-x-[GF]-x-C-x(5)-C-x(3)-C [The four C's are involved in disulfide bonds]-

- 5 [1] Bruix M., Jimenez M.A., Santoro J., Gonzalez C., Colilla F.J., Mendez E., Rico M. Biochemistry 32:715-724(1993).
- [2] Gu Q., Kawata E.E., Morse M.-J., Wu H.-M., Cheung A.Y. Mol. Gen. Genet. 234:89-96(1992).
- [3] Terras F.R.G., Torrekens S., van Leuven F., Osborn R.W., Vanderleyden J., Cammue
10 B.P.A., Broekaert W.F. FEBS Lett. 316:233-240(1993).
- [4] Bloch C. Jr., Richardson M. FEBS Lett. 279:101-104(1991).
- [5] Ishibashi N., Yamauchi D., Miniamikawa T. Plant Mol. Biol. 15:59-64(1990).
- [7] Choi Y., Choi Y.D., Lee J.S. Plant Physiol. 101:699-700(1993).

231. Gelsolin. Gelsolin repeat. Number of members: 170

[1]Medline: 97433077. The crystal structure of plasma gelsolin: implications for actin severing, capping, and nucleation. Burtnick LD, Koepf EK, Grimes J, Jones EY, Stuart DI, McLaughlin PJ, Robinson RC; Cell 1997;90:661-670.

232. Germin family signature

Germins [1] are a family of homopentameric cereal glycoproteins expressed during
25 germination which may play a role in altering the properties of cell walls during germinative growth. It has been shown that wheat and barleygermins act as oxalate oxidases (EC 1.2.3.4), an enzyme that catalyzes the oxidative degradation of oxalate to carbonate and hydrogen peroxide. Germins are highly similar to: - Germin-like proteins from various plants such as rape, violet or white mustard. - Slime mold spherulins 1a and 1b which are proteins that
30 accumulate specifically during spherulation, a process induced by various forms of environmental stress which leads to encystment and dormancy. As a signature pattern the best conserved region was selected: a decapeptide located in the central section of these proteins.

Consensus pattern: G-x(4)-H-x-H-P-x-A-x-E-[LIVM]-

[1] Lane B.G. FASEB J. 8:294-301(1994).

5

233. (GluTR)

Glutamyl-tRNA reductase signature

10

Delta-aminolevulinic acid (ALA) is the obligatory precursor for the synthesis of all tetrapyrroles including porphyrin derivatives such as chlorophyll and heme. ALA can be synthesized via two different pathways: the Shemin (or C4) pathway which involves the single step condensation of succinyl-CoA and glycine and which is catalyzed by ALA synthase (EC 2.3.1.37) and via the C5 pathway from the five-carbon skeleton of glutamate. The C5 pathway operates in the chloroplast of plants and algae, in cyanobacteria, in some eubacteria and in archaeobacteria.

15

The initial step in the C5 pathway is carried out by glutamyl-tRNA reductase (GluTR) [1] which catalyzes the NADP-dependent conversion of glutamate- tRNA(Glu) to glutamate-1-semialdehyde (GSA) with the concomitant release of tRNA(Glu) which can then be recharged with glutamate by glutamyl-tRNA synthetase.

20

GluTR is a protein of about 50 Kd (467 to 550 residues) which contains a few conserved region. The best conserved region is located in positions 99 to 122 in the sequence of known GluTR. This region seems important for the activity of the enzyme. We have developed a signature pattern from that conserved region.

25

Consensus pattern H-[LIVM]-x(2)-[LIVM]-[GSTAC](3)-[LIVM]-[DEQ]-S-[LIVMA]-[LIVM](2)-[GF]-E-x-[EQR]-[IV]-[LIT]-[STAG]-Q-[LIVM]-[KR] Sequences known to belong to this class detected by the pattern ALL.

30

[1] Jahn D., Verkamp E., Soell D. Trends Biochem. Sci. 17:215-218(1992).

234. (Glycoprotease)

Glycoprotease family signature (aka Peptidase_M22)

Glycoprotease (GCP) (EC 3.4.24.57) [1], or o-sialoglycoprotein endopeptidase, is a metalloprotease secreted by *Pasteurella haemolytica* which specifically cleaves O-sialoglycoproteins such as glycophorin A. The sequence of GCP is highly similar to the following uncharacterized proteins:

- *Escherichia coli* hypothetical protein ygjD (ORF-X).
- *Bacillus subtilis* hypothetical protein ydiE.
- *Mycobacterium leprae* hypothetical protein U229E.
- *Mycobacterium tuberculosis* hypothetical protein MtCY78.10.
- *Synechocystis* strain PCC 6803 hypothetical protein slr0807.
- *Methanococcus jannaschii* hypothetical protein MJ1130.
- *Haloarcula marismortui* hypothetical protein in HSH 3' region.
- Yeast hypothetical protein YKR038c.
- Yeast hypothetical protein QRI7.

One of the conserved regions contains two conserved histidines. It is possible that this region is involved in coordinating a metal ion such as zinc.

Consensus pattern[KR]-[GSAT]-x(4)-[FYWLH]-[DQNGK]-x-P-x-[LIVMFY]-x(3)-H- x(2)-[AG]-H-[LIVM] Sequences known to belong to this class detected by the pattern ALL.

Note: these proteins belong to family M22 in the classification of peptidases [2,E1].

□

[1] Abdullah K.M., Lo R.Y.C., Mellors A. J. Bacteriol. 173:5597-5603(1991).

□

[2] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).

235. (Glucosamine_iso)

260

Glucosamine/galactosamine-6-phosphate isomerases signature

Glucosamine-6-phosphate isomerase (EC 5.3.1.10) (or Glc-6-P deaminase) is the enzyme responsible for the conversion of glucosamine 6-phosphate into fructose6 phosphate [1]. It is the last specific step in the pathway for N-acetylglucosamine (GlcNAC) utilization in bacteria such as *Escherichia coli* (gene nagB) or in fungi such as *Candida albicans* (gene NAG1). Glc-6-P isomerase is evolutionary related to: - A putative *Escherichia coli* galactosamine-6-phosphate isomerase (gene agaI) [2]. - *Escherichia coli* hypothetical protein yieK. - *Bacillus subtilis* hypothetical protein ybfT. As a signature pattern a conserved region located in the central part of these enzymes was selected. This region contains a conserved histidine which has been shown [1], in nagB, to be important for the pyranose ring-opening step of the catalytic mechanism

Consensus pattern: [LIVM]-x(3)-G-x-[LIT]-x-[LIV]-x-[LIVM]-x-G-[LIVM]-G-x- [DEN]-G-H-

[1] Oliva G., Fontes M.R.M., Garratt R.C., Altamirano M.M., Calcagno M.L., Horjaes E. Structure 3:1323-1332(1995).

[2] Reizer J., Ramseier T.M., Reizer A., Charbit A., Saier M.H. Jr. Microbiology 142:231-250(1996).

236. Pneumovirus attachment glycoprotein G (glycoprotein G)

This family includes attachment proteins from respiratory syncytial virus. Glycoprotein G has not been shown to have any neuraminidase or hemagglutinin activity (Swiss-Prot). The amino terminus is thought to be cytoplasmic, and the carboxyl terminus extracellular. The extracellular region contains four completely conserved cysteine residues.

[1] Johnson PR, Spriggs MK, Olmsted RA, Collins PL, Proc Natl Acad Sci U S A 1987;84:5625-5629.

237. Glycosyl transferases group 1

Mutations in this domain of Swiss:P37287 lead to disease (Paroxysmal Nocturnal haemoglobinuria). Members of this family transfer activated sugars to a variety of substrates,

including glycogen, Fructose-6-phosphate and lipopolysaccharides. Members of this family transfer UDP, ADP, GDP or CMP linked sugars. The eukaryotic glycogen synthases may be distant members of this family.

5

238. Glycosyl transferases (Glycos_transf_2)

Diverse family, transferring sugar from UDP-glucose, UDP-N-acetyl-galactosamine, GDP-mannose or CDP-abequose, to a range of substrates including cellulose, dolichol phosphate and teichoic acids.

10

239. (Glucos_transf_3)

Thymidine and pyrimidine-nucleoside phosphorylases signature

15

Thymidine phosphorylase (EC 2.4.2.4) catalyzes the reversible phosphorolysis of thymidine, deoxyuridine and their analogues to their respective bases and 2-deoxyribose 1-phosphate. This enzyme regulates the availability of thymidine and is therefore essential to nucleic acid metabolism.

20

In *Escherichia coli* (gene *deoA*), the enzyme is a dimer of identical subunits of about 48 Kd [1]. In humans it was first identified as platelet-derived endothelial cell growth factor (PD-ECGF) [E1] before being recognized [2] as thymidine phosphorylase.

25

Bacterial pyrimidine-nucleoside phosphorylase (EC 2.4.2.2) (gene *pdp*) [3] is an enzyme evolutionary and structurally related to thymidine phosphorylase.

A well conserved region of 19 residues located in the N-terminal part of these proteins signature pattern for these enzymes was selected.

30

Consensus pattern S-[GS]-R-[GA]-[LIV]-x(2)-[TA]-[GA]-G-T-x-D-x-[LIV]-E Sequences known to belong to this class detected by the pattern ALL.

[1] Walter M.R., Cook W.J., Cole L.B., Short S.A., Koszalka G.W., Krenitsky T.A., Ealick S.E. J. Biol. Chem. 265:14016-14022(1990).

[2] Furukawa T., Yoshimura A., Sumizawa T., Haraguchi M., Akiyama S.-I., Fukui K., Yamada Y. Nature 356:668-668(1992).

5 [3] Saxild H.H., Andersen L.N., Hammer K. J. Bacteriol. 178:424-434(1996).

240. Glycos_transf_4. Glycosyl transferase. Number of members: 44.

10 [1] Medline: 95252686. A family of UDP-GlcNAc/MurNAc: polyisoprenol-P GlcNAc/MurNAc-1-P transferases. Lehrman MA; Glycobiology 1994;4:768-771.

241. Glycosyl hydrolases family 15. 21 members.

242. Glycosyl hydrolases family 16 signature

It has been shown [1] that the following glycosyl hydrolases can be classified into a single family on the basis of sequence similarities: - Bacterial beta-1,3-1,4-glucanases, or lichenases, (EC 3.2.1.73) mainly from *Bacillus* but also from *Clostridium thermocellum* (gene licB), *Fibrobacter succinogenes* and *Rhodothermus marinus* (gene bglA). - *Bacillus circulans* beta-1,3-glucanase A1 (EC 3.2.1.39) (gene glcA). - Lamarinase (EC 3.2.1.6) from *Clostridium thermocellum* (gene lam1). - *Streptomyces coelicolor* agarase (EC 3.2.1.81) (gene dagA). - *Alteromonas carrageenovora* kappa-carrageenase (EC 3.2.1.83) (gene cgkA). Two closely clustered conserved glutamates have been shown [2] to be involved in the catalytic activity of *Bacillus licheniformis* lichenase. The region was used that contains these residues as a signature pattern.

Consensus pattern: E-[LIV]-D-[LIV]-x(0,1)-E-x(2)-[GQ]-[KRNF]-x-[PSTA] [The two E's are active site residues]-

[1] Henrissat B. Biochem. J. 280:309-316(1991).

[2] Juncosa M., Pons J., Dot T., Querol E., Planas A. J. Biol. Chem. 269:14530-14535(1994).

5 243. Glycosyl hydrolases family 17 signature

It has been shown [1,2] that the following glycosyl hydrolases can be classified into a single family on the basis of sequence similarities: - Glucan endo-1,3-beta-glucosidases (EC 3.2.1.39) (endo-(1->3)-beta- glucanase) from various plants. This enzyme may be involved in the defense of plants against pathogens through its ability to degrade fungal cell wall polysaccharides. - Glucan 1,3-beta-glucosidase (EC 3.2.1.58) (exo-(1->3)-beta-glucanase) from yeast (gene BGL2). This enzyme may play a role in cell expansion during growth, in cell-cell fusion during mating, and in spore release during sporulation. - Lichenases (EC 3.2.1.73) (endo-(1->3,1->4)-beta-glucanase) from various plants. The best conserved region in the sequence of these enzymes is located in their central section. This region contains a conserved tryptophan residue which could be involved in the interaction with the glucan substrates [2] and it also contains a conserved glutamate which has been shown [3] to act as the nucleophile in the catalytic mechanism. this region was used as a signature pattern.

Consensus pattern: [LIVM]-x-[LIVMFYWA](3)-[STAG]-E-[STA]-G-W-P-[STN]-x-[SAGQ] [E is an active site residue]-

[1] Henrissat B. Biochem. J. 280:309-316(1991).

[2] Ori N., Sessa G., Lotan T., Himmelhoch S., Fluhr R. EMBO J. 9:3429-3436(1990).

[3] Varghese J.N., Garrett T.P.J., Colman P.M., Chen L., Hoj P.J., Fincher G.B. Proc. Natl. Acad. Sci. U.S.A. 91:2785-2789(1994).

244. Glyoxalase I signatures

Glyoxalase I (EC 4.4.1.5) (lactoylglutathione lyase) catalyzes the first step of the glyoxal pathway, the transformation of methylglyoxal and glutathione into S-lactoylglutathione which is then converted by glyoxalase II to lactic acid [1]. Glyoxalase I is an ubiquitous enzyme which binds one mole of zinc per subunit. The bacterial and yeast enzymes are monomeric while the mammalian one is homodimeric. The sequence of glyoxalase I is well conserved. In

bacteria and mammals, the enzyme is a protein of about 130 to 180 residues while in fungi it is about twice longer. In these organisms the enzyme is built out of the tandem repeat of an homologous domain. Two signature patterns for this family were derived. The first one is located in the N-terminal region while the second one is located in the central section of the protein and contains a conserved histidine that could be implicated in the binding of the zinc atom.

Consensus pattern: [HQ]-[IVT]-x-[LIVFY]-x-[IV]-x(5)-[STA]-x(2)-F-[YM]-x(2,3)-[LMF]-G-[LMF]-

Consensus pattern: G-[NTKQ]-x(0,5)-[GA]-[LVFY]-[GH]-H-[IVF]-[CGA]-x-[STAGLE]-x(2)-[DNC]-

[1] Kim N.-S., Umezawa Y., Ohmura S., Kato S. J. Biol. Chem. 268:11217-11221(1993).

245. (Glypican)

Glypicans signature

Glypicans [1,2] are a family of heparan sulfate proteoglycans which are anchored to cell membranes by a glycosylphosphatidylinositol (GPI) linkage. Structurally, these proteins consist of three separate domains:

- a) A signal sequence;
- b) An extracellular domain of about 500 residues that contains 12 conserved cysteines probably involved in disulfide bonds and which also contains the sites of attachment of the heparan sulfate glycosaminoglycan side chains;
- c) A C-terminal hydrophobic region which is post-translationally removed after formation of the GPI-anchor.

The proteins known to belong to this family are:

- Glypican 1 (GPC1).
- Glypican 2 (GPC2) or cerebroglycan.

- Glypican 3 (GPC3) or OCI-5. In man, defects in GPC3 are the cause of a X-linked genetic disease, Simpson-Galabi-Behmel syndrome (SGBS).

- K-glypican.

- Glypican 5 (GPC5).

5 - Drosophila protein dally.

The signature pattern that was developed for glypicans is located in the central section of the extracellular domain and contains five of the conserved cysteines.

10 Consensus pattern C-x(2)-C-x-G-[LIVM]-x(4)-P-C-x(2)-[FY]-C-x(2)-[LIVM]-x(2)-G-C [The C's are probably involved in a disulfide bonds] Sequences known to belong to this class detected by the pattern ALL, except for dally.

[1] Weksberg R., Squire J.A., Templeton D.M. Nat. Genet. 12:225-227(1996).

15 [2] Watanabe K., Yamada H., Yamaguchi Y. J. Cell Biol. 130:1207-1218(1995).

246. Granins signatures

Granins (chromogranins or secretogranins) [1] are a family of acidic proteins present in the secretory granules of a wide variety of endocrine and neuro-endocrine cells. The exact function(s) of these proteins is not yet known but they seem to be the precursors of biologically active peptides and/or they may act as helper proteins in the packaging of peptide hormones and neuropeptides. Three members of this family of proteins show some sequence similarities: - Chromogranin A (CGA) [2]. CGA is a protein of about 420 residues; it is the precursor of the peptide pancreastatin which strongly inhibits glucose- induced insulin release from the pancreas. - Secretogranin 1 (chromogranin B). A sulfated protein of about 600 residues. - Secretogranin 2 (chromogranin C). A sulfated protein of about 650 residues. Apart from their subcellular location and the abundance of acidic residues(Asp and Glu), these proteins do not share many structural similarities. Only one short region, located in the C-terminal section, is conserved in all these proteins. Chromogranins A and B share a region of high similarity in their N-terminal section; this region includes two cysteine residues involved in a disulfide bond

Consensus pattern: [DE]-[SN]-L-[SAN]-x(2)-[DE]-x-E-L-

Consensus pattern: C-[LIVM](2)-E-[LIVM](2)-S-[DN]-[STA]-L-x-K-x-S-x(3)- [LIVM]-[STA]-x-E-C [The two C's are linked by a disulfide bond]-

5 [1] Huttner W.B., Gerdes H.-H., Rosa P. Trends Biochem. Sci. 16:27-30(1991).

[2] Simon J.-P., Aunis D. Biochem. J. 262:1-13(1989).

247. grpE protein signature

10 In prokaryotes the grpE protein [1] stimulates, jointly with dnaJ, the ATPase activity of the dnaK chaperone. It seems to accelerate the release of ADP from dnaK thus allowing dnaK to recycle more efficiently. GrpE is a protein of about 22 to 25 Kd. In yeast, an evolutionary related mitochondrial protein(gene GRPE) has been shown [2] to associate with the mitochondrial hsp70protein and to thus play a role in the import of proteins from the cytoplasm. As a signature pattern, the most conserved region of grpE was selected. It is located in the C-terminal section.

Consensus pattern: [FL]-[DN]-[PHEA]-x(2)-[HM]-x-A-[LIVMTN]-x(16,20)-G-[FY]- x(3)-[DEG]-x(2)-[LIVM]-[RI]-x-[SA]-x-V-x-[IV]-

20 [1] Georgopoulos C., Welch W. Annu. Rev. Cell Biol. 9:601-635(1993).

[2] Bolliger L., Deloche O., Glick B.S., Georgopoulos C., Jenoe P., Kronidou N., Horst M., Morishima N., Schatz G. EMBO J. 13:1998-2006(1994).

25

248. Guanylate kinase signature and profile

30 Guanylate kinase (EC 2.7.4.8) (GK) [1] catalyzes the ATP-dependent phosphorylation of GMP into GDP. It is essential for recycling GMP and indirectly, cGMP. In prokaryotes (such as Escherichia coli), lower eukaryotes (such as yeast) and in vertebrates, GK is a highly conserved monomeric protein of about 200 amino acids. GK has been shown [2,3,4] to be structurally similar to the following proteins: - Protein A57R (or SalG2R) from various strains of Vaccinia virus. This protein is highly similar to GK, but contains a frameshift mutation in the N-terminal section and could therefore be inactive in that virus. The

following proteins are characterized by the presence in their sequence of one or more copies of the DHR domain, a SH3 domain (see <PDOC50002> as well as a C-terminal GK-like domain, these proteins are collectively termed MAGUKs (membrane-associated guanylate kinase homologs) [5]: - *Drosophila* lethal(1) discs large-1 tumor suppressor protein (gene *dlg1*). This protein is associated with septate junctions in developing flies and defects in the *dlg1* gene cause neoplastic overgrowth of the imaginal disks. - Mammalian tight junction protein Zo-1. - A family of mammalian synaptic proteins that seem to interact with the cytoplasmic tail of NMDA receptor subunits. This family currently consists of SAP90/PSD-95, CHAPSYN-110/PSD-93, SAP97/DLG1 and SAP102. - Vertebrate 55 Kd erythrocyte membrane protein (p55). p55 is a palmitoylated, membrane-associated protein of unknown function. - *Caenorhabditis elegans* protein lin-2, which may play a structural role in the induction of the vulva. - Rat protein CASK. - Human protein DLG2. - Human protein DLG3. There is an ATP-binding site (P-loop) in the N-terminal section of GK. This region is not conserved in the GK-like domain of the above proteins which are therefore unlikely to be kinases. However these proteins retain the residues known, in GK, to be involved in the binding of GMP. As a signature pattern a highly conserved region was selected that contains two arginine and a tyrosine which are involved in GMP-binding

Consensus pattern: T-[ST]-R-x(2)-[KR]-x(2)-[DE]-x(2)-G-x(2)-Y-x-[FY]-[LIVMK]-

[1] Stehle T., Schulz G.E. *J. Mol. Biol.* 224:1127-1141(1992).

[2] Bryant P.J., Woods D.F. *Cell* 68:621-622(1992).

[3] Goebl M.G. *Trends Biochem. Sci.* 17:99-99(1992).

[4] Zschocke P.D., Schiltz E., Schulz G.E. *Eur. J. Biochem.* 213:263-269(1993).

[5] Woods D.F., Bryant P.J. *Mech. Dev.* 44:85-89(1994).

249. (Glyco_hydro_35)

Glycosyl hydrolases family 35 putative active site

Beta-galactosidases (EC 3.2.1.23) from mammals, fungi, plants and the bacteria

Xanthomonas manihotis are evolutionary related [1,2]. They belong to family 35 in the classification of glycosyl hydrolases [3,E1].

Mammalian beta-galactosidase is a lysosomal enzyme (gene GLB1) which cleaves the terminal galactose from gangliosides, glycoproteins, and glycosaminoglycans and whose deficiency is the cause of the genetic disease Gm(1) gangliosidosis (Morquio disease type B).

5

On of the best conserved regions in these enzymes contains a glutamic acid residue which, on the basis of similarities with other families of glycosyl hydrolases [4], probably acts as the proton donor in the catalytic mechanism. This region wss used as a signature pattern.

10 Consensus pattern: G-G-P-[LIVM](2)-x(2)-Q-x-E-N-E-[FY] [The second E is the putative active site residue] Sequences known to belong to this class detected by the pattern ALL.

[1] Taron C.H., Benner J.S., Hornstra L.J., Guthrie E.P. Glycobiology 5:603-610(1995).

[2] Carey A.T., Holt K., Picard S., Wilde R., Tucker G.A., Bird C.R., Schuch W., Seymour G.B. Plant Physiol. 108:1099-1107(1995).

[3] Henrissat B., Bairoch A. Biochem. J. 293:781-788(1993).

[4] Henrissat B., Callebaut I., Fabrega S., Lehn P., Mornon J.-P., Davies G. Proc. Natl. Acad. Sci. U.S.A. 92:7090-7094(1995).

20

250. (Glyco_hydro_16)

Glycosyl hydrolases family 16 signature

It has been shown [1] that the following glycosyl hydrolases can be classified into a single family on the basis of sequence similarities:

25

- Bacterial beta-1,3-1,4-glucanases, or lichenases, (EC 3.2.1.73) mainly from *Bacillus* but also from *Clostridium thermocellum* (gene *licB*), *Fibrobacter succinogenes* and *Rhodothermus marinus* (gene *bglA*).
- 30 - *Bacillus circulans* beta-1,3-glucanase A1 (EC 3.2.1.39) (gene *glcA*).
- Laminase (EC 3.2.1.6) from *Clostridium thermocellum* (gene *lam1*).
- *Streptomyces coelicolor* agarase (EC 3.2.1.81) (gene *dagA*).
- *Alteromonas carrageenovora* kappa-carrageenase (EC 3.2.1.83) (gene *cgkA*).

Two closely clustered conserved glutamates have been shown [2] to be involved in the catalytic activity of *Bacillus licheniformis* lichenase. The region that contains these residues as a signature pattern was used.

5

Consensus pattern E-[LIV]-D-[LIV]-x(0,1)-E-x(2)-[GQ]-[KRNF]-x-[PSTA] [The two E's are active site residues]

[1] Henrissat B. Biochem. J. 280:309-316(1991).

10 [2] Juncosa M., Pons J., Dot T., Querol E., Planas A. J. Biol. Chem. 269:14530-14535(1994).

251. (Glyco_hydro_17)

15

Glycosyl hydrolases family 17 signature
(aka glycosyl_hydro4)

It has been shown [1,2] that the following glycosyl hydrolases can be classified into a single family on the basis of sequence similarities:

20

- Glucan endo-1,3-beta-glucosidases (EC 3.2.1.39) (endo-(1->3)-beta-glucanase) from various plants. This enzyme may be involved in the defense of plants against pathogens through its ability to degrade fungal cell wall polysaccharides.

25

- Glucan 1,3-beta-glucosidase (EC 3.2.1.58) (exo-(1->3)-beta-glucanase) from yeast (gene BGL2). This enzyme may play a role in cell expansion during growth, in cell-cell fusion during mating, and in spore release during sporulation.

- Lichenases (EC 3.2.1.73) (endo-(1->3,1->4)-beta-glucanase) from various plants.

The best conserved region in the sequence of these enzymes is located in their central section.

30 This region contains a conserved tryptophan residue which could be involved in the interaction with the glucan substrates [2] and it also contains a conserved glutamate which has been shown [3] to act as the nucleophile in the catalytic mechanism. This region was used as a signature pattern.

Consensus pattern [LIVM]-x-[LIVMFYWA](3)-[STAG]-E-[STA]-G-W-P-[STN]-x-[SAGQ]
 [E is an active site residue] Sequences known to belong to this class detected by the pattern
 ALL.

5

- [1] Henrissat B. Biochem. J. 280:309-316(1991).
- [2] Ori N., Sessa G., Lotan T., Himmelhoch S., Fluhr R. EMBO J. 9:3429-3436(1990).
- [3] Varghese J.N., Garrett T.P.J., Colman P.M., Chen L., Hoj P.J., Fincher G.B. Proc. Natl.
 Acad. Sci. U.S.A. 91:2785-2789(1994).

10

252. (Glyco_hydro_3)

Glycosyl hydrolases family 3 active site

It has been shown [1,2] that the following glycosyl hydrolases can be, on the basis of
 sequence similarities, classified into a single family:

- Beta glucosidases (EC 3.2.1.21) from the fungi *Aspergillus wentii* (A-3),
Hansenula anomala, *Kluyveromyces fragilis*, *Saccharomycopsis fibuligera*,
 (BGL1 and BGL2), *Schizophyllum commune* and *Trichoderma reesei* (BGL1).
- Beta glucosidases from the bacteria *Agrobacterium tumefaciens* (Cbg1),
Butyrivibrio fibrisolvens (bglA), *Clostridium thermocellum* (bglB),
Escherichia coli (bglX), *Erwinia chrysanthemi* (bgxA) and *Ruminococcus*
albus.
- *Alteromonas* strain O-7 beta-hexosaminidase A (EC 3.2.1.52).
- *Bacillus subtilis* hypothetical protein yzbA.
- *Escherichia coli* hypothetical protein ycfO and HI0959, the corresponding
Haemophilus influenzae protein.

25

30 One of the conserved regions in these enzymes is centered on a conserved aspartic acid
 residue which has been shown [3], in *Aspergillus wentii* beta- glucosidase A3, to be
 implicated in the catalytic mechanism. This region was used as a signature pattern.

271

Consensus pattern[LIVM](2)-[KR]-x-[EQK]-x(4)-G-[LIVMFT]-[LIVT]-[LIVMF]- [ST]-D-x(2)-[SGADNI] [D is the active site residue] Sequences known to belong to this class detected by the patternALL.

- 5 [1] Henrissat B. Biochem. J. 280:309-316(1991).
 [2] Castle L.A., Smith K.D., Morris R.O. J. Bacteriol. 174:1478-1486(1992).
 [3] Bause E., Legler G. Biochim. Biophys. Acta 626:459-465(1980).

10 253. (Glyco_hydro_28)
 Polygalacturonase active site (aka PG)

Polygalacturonase (EC 3.2.1.15) (PG) (pectinase) [1,2] catalyzes the random hydrolysis of 1,4-alpha-D-galactosiduronic linkages in pectate and other galacturonans. In fruit, polygalacturonase plays an important role in cell wall metabolism during ripening. In plant bacterial pathogens such as *Erwinia carotovora* or *Pseudomonas solanacearum* and fungal pathogens such as *Aspergillus niger*, polygalacturonase is involved in maceration and soft-rotting of plant tissue.

20 Exo-poly-alpha-D-galacturonosidase (EC 3.2.1.82) (exoPG) [3] hydrolyzes peptic acid from the non-reducing end, releasing digalacturonate.

25 Prokaryotic, eukaryotic PG and exoPG share a few regions of sequence similarity. The best conserved of these regions was selected. It is centered on a conserved histidine most probably involved in the catalytic mechanism [4].

Consensus pattern[GSDENKRH]-x(2)-[VMFC]-x(2)-[GS]-H-G-[LIVMAG]-x(1,2)- [LIVM]-G-S [H is the putative active site residue] Sequences known to belong to this class detected by the patternALL.

30 Note: these proteins belong to family 28 in the classification of glycosyl hydrolases [5].

[1] Ruttowski E., Labitzke R., Khanh N.Q., Loeffler F., Gottschalk M., Jany K.-D. Biochim. Biophys. Acta 1087:104-106(1990).

[2] Huang J., Schell M.A. J. Bacteriol. 172:3879-3887(1990).

[3] He S.Y., Collmer A. J. Bacteriol. 172:4988-4995(1990).

5 [4] Bussink H.J.D., Buxton F.P., Visser J. Curr. Genet. 19:467-474(1991).

[5] Henrissat B. Biochem. J. 280:309-316(1991).

254. (Glyco_hydro_32)

10 Glycosyl hydrolases family 32 active site

It has been shown [1,2] that the following glycosyl hydrolases can be classified into a single family on the basis of sequence similarities:

- 15 - Inulinase (EC 3.2.1.7) (or inulase) from the fungi *Kluyveromyces marxianus*.
- Beta-fructofuranosidase (EC 3.2.1.26), commonly known as invertase in fungi and plants and as sucrase in bacteria (gene *sacA* or *scrB*).
- Raffinose invertase (EC 3.2.1.26) (gene *rafD*) from *Escherichia coli* plasmid pRSD2.
- 20 - Levanase (EC 3.2.1.65) (gene *sacC*) from *Bacillus subtilis*.

One of the conserved regions in these enzymes is located in the N-terminal section and contains an aspartic acid residue which has been shown [3], in yeast invertase to be important for the catalytic mechanism. This region was used as a signature pattern.

25

Consensus pattern H-x(2)-P-x(4)-[LIVM]-N-D-P-N-G [D is the active site residue]

Sequences known to belong to this class detected by the patternALL.

[1] Henrissat B. Biochem. J. 280:309-316(1991).

30 [2] Gunasekaran P., Karunakaran T., Cami B., Mukundan A.G., Preziosi L., Baratti J. J. Bacteriol. 172:6727-6735(1990).

[3] Reddy V.A., Maley F. J. Biol. Chem. 265:10817-10120(1990).

255. (Glyco_hydro_1)

Glycosyl hydrolases family 1 signatures

5 It has been shown [1 to 4] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:

- Beta-glucosidases (EC 3.2.1.21) from various bacteria such as *Agrobacterium* strain ATCC 21400, *Bacillus polymyxa*, and *Caldocellum saccharolyticum*.

10 - Two plants (clover) beta-glucosidases (EC 3.2.1.21).

- Two different beta-galactosidases (EC 3.2.1.23) from the archaebacteria *Sulfolobus solfataricus* (genes *bgaS* and *lacS*).

- 6-phospho-beta-galactosidases (EC 3.2.1.85) from various bacteria such as *Lactobacillus casei*, *Lactococcus lactis*, and *Staphylococcus aureus*.

15 - 6-phospho-beta-glucosidases (EC 3.2.1.86) from *Escherichia coli* (genes *bglB* and *ascB*) and from *Erwinia chrysanthemi* (gene *arbB*).

- Plants myrosinases (EC 3.2.3.1) (sinigrinase) (thioglucosidase).

- Mammalian lactase-phlorizin hydrolase (LPH) (EC 3.2.1.108 / EC 3.2.1.62).

20 LPH, an integral membrane glycoprotein, is the enzyme that splits lactose in the small intestine. LPH is a large protein of about 1900 residues which contains four tandem repeats of a domain of about 450 residues which is evolutionary related to the above glycosyl hydrolases.

25 One of the conserved regions in these enzymes is centered on a conserved glutamic acid residue which has been shown [5], in the beta-glucosidase from *Agrobacterium*, to be directly involved in glycosidic bond cleavage by acting as a nucleophile. This region was used as a signature pattern. As a second signature pattern we selected a conserved region, found in the N-terminal extremity of these enzymes, this region also contains a glutamic acid residue.

30 Consensus pattern[LIVMFSTC]-[LIVFYS]-[LIV]-[LIVMST]-E-N-G-[LIVMFAR]-[CSAGN] [E is the active site residue] Sequences known to belong to this class detected by the patternALL.

274

Note: this pattern will pick up the last two domains of LPH; the first two domains, which are removed from the LPH precursor by proteolytic processing, have lost the active site glutamate and may therefore be inactive [4].

- 5 Consensus pattern F-x-[FYWM]-[GSTA]-x-[GSTA]-x-[GSTA](2)-[FYNH]-[NQ]-x-E-x-[GSTA] Sequences known to belong to this class detected by the pattern ALL.

Note: this pattern will pick up the last three domains of LPH.

- 10 [1] Henrissat B. Biochem. J. 280:309-316(1991).
 [2] Henrissat B. Protein Seq. Data Anal. 4:61-62(1991).
 [3] Gonzalez-Candelas L., Ramon D., Polaina J. Gene 95:31-38(1990).
 [4] El Hassouni M., Henrissat B., Chippaux M., Barras F. J. Bacteriol. 174:765-777(1992).
 [5] Withers S.G., Warren R.A.J., Street I.P., Rupitz K., Kempton J.B., Aebersold R. J. Am. Chem. Soc. 112:5887-5889(1990).

256. Glyco_hydro_20

Glycosyl hydrolase family 20

Previous Pfam IDs: glycosyl_hydr11;

Number of members: 33

257. (Glyco_hydro_9)

- 25 Glycosyl hydrolases family 9 active sites signatures
 (aka Glycosyl_hydr12)

- 30 The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1,2]. Fungi and bacteria produces a spectrum of cellulolytic enzymes (cellulases) and xylanases which, on the basis of sequence similarities, can be classified into families. One of these families is known as the cellulase family E [3] or as the glycosyl

hydrolases family 9 [4,E1]. The enzymes which are currently known to belong to this family are listed below.

- Butyrivibrio fibrisolvens cellodextrinase 1 (ced1).
- Cellulomonas fimi endoglucanases B (cenB) and C (cenC).
- Clostridium cellulolyticum endoglucanase G (celCCG).
- Clostridium cellulovorans endoglucanase C (engC).
- Clostridium stercoararium endoglucanase Z (avicelase I) (celZ).
- Clostridium thermocellum endoglucanases D (celD), F (celF) and I (celI).
- Fibrobacter succinogenes endoglucanase A (endA).
- Pseudomonas fluorescens endoglucanase A (celA).
- Streptomyces reticuli endoglucanase 1 (cel1).
- Thermomonospora fusca endoglucanase E-4 (celD).
- Dictyostelium discoideum spore germination specific endoglucanase 270-6. This slime mold enzyme may digest the spore cell wall during germination, to release the enclosed amoeba.
- Endoglucanases from plants such as Avocado or French bean. In plants this enzyme may be involved the fruit ripening process.

Two of the most conserved regions in these enzymes are centered on conserved residues which have been shown [5,6], in the endoglucanase D from Cellulomonas thermocellum, to be important for the catalytic activity. The first region contains an active site histidine and the second region contains two catalytically important residues: an aspartate and a glutamate.

Both regions were used as signature patterns.

Consensus pattern [STV]-x-[LIVMFY]-[STV]-x(2)-G-x-[NKR]-x(4)-[PLIVM]-H-x-R [H is an active site residue] Sequences known to belong to this class detected by the pattern ALL, except for Cellulomonas fimi cenC and Streptomyces reticuli cel1.

Consensus pattern [FYW]-x-D-x(4)-[FYW]-x(3)-E-x-[STA]-x(3)-N-[STA] [D and E are active site residues] Sequences known to belong to this class detected by the pattern ALL, except for Fibrobacter succinogenes endA whose sequence seems to be incorrect.

- [1] Beguin P. Annu. Rev. Microbiol. 44:219-248(1990).
- [2] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).
- 5 [3] Henrissat B., Claeysens M., Tomme P., Lemesle L., Mornon J.-P. Gene 81:83-95(1989).
- [4] Henrissat B. Biochem. J. 280:309-316(1991).
- [5] Tomme P., Chauvaux S., Beguin P., Millet J., Aubert J.-P., Claeysens M. J. Biol. Chem. 266:10313-10318(1991).
- [6] Tomme P., van Beeumen J., Claeysens M. Biochem. J. 285:319-324(1992).

10

258. Matrix protein (MA), p15 (GAG_ma).

The matrix protein, p15, is encoded by the gag gene. MA is involved in pathogenicity [1].

15 [1] : Pozsgay JM, Beilharz MW, Wines BD, Hess AD, Pitha PM, J Virol 1993;67:5989-5999.

259. Gag polyprotein, inner coat protein p12 (GAG_P12)

20 The retroviral p12 is a virion structural protein. p12 is proline rich. The function carried out by p12 in assembly and replication is unknown. p12C is associated with pathogenicity of the virus

[1] Pozsgay JM, Beilharz MW, Wines BD, Hess AD, Pitha PM, J Virol 1993;67:5989-5999.

25

260. Glutamine synthetase signatures (GLN-SYNT)

Glutamine synthetase (EC 6.3.1.2) (GS) [1] plays an essential role in the metabolism of nitrogen by catalyzing the condensation of glutamate and ammonia to form glutamine. There seem to be three different classes of GS [2,3,4]: - Class I enzymes (GSI) are specific to prokaryotes, and are oligomers of 12 identical subunits. The activity of GSI-type enzyme is controlled by the adenylation of a tyrosine residue. The adenylation of the enzyme is inactive. - Class II enzymes (GSII) are found in eukaryotes and in bacteria belonging to the Rhizobiaceae, Frankiaceae, and Streptomycetaceae families (these bacteria have also a class-I

30

GS). GSII are octamer of identical subunits. Plants have two or more isozymes of GSII, one of the isozymes is translocated into the chloroplast. - Class III enzymes (GSIII) has, currently, only been found in *Bacteroides fragilis* and in *butyrivibrio fibrisolvens*. It is a hexamer of identical chains. It is much larger (about 700 amino acids) than the GSI (450 to 470 amino acids) or GSII (350 to 420 amino acids) enzymes. While the three classes of GS's are clearly structurally related, the sequence similarities are not so extensive. As signature patterns three conserved regions were selected. The first pattern is based on a conserved tetrapeptide in the N-terminal section of the enzyme, the second one is based on a glycine-rich region which is thought to be involved in ATP-binding. The third pattern is specific to class I glutamine synthetases and includes the tyrosine residue which is reversibly adenylated.

Consensus pattern: [FYWL]-D-G-S-S-x(6,8)-[DENQSTAK]-[SA]-[DE]-x(2)-[LIVMFY]-

Consensus pattern: K-P-[LIVMFYA]-x(3,5)-[NPAT]-G-[GSTAN]-G-x-H-x(3)-S-

Consensus pattern: K-[LIVM]-x(5)-[LIVMA]-D-[RK]-[DN]-[LI]-Y [Y is the site of adenylation]-

[1] Eisenberg D., Almassy R.J., Janson C.A., Chapman M.S., Suh S.W., Cascio D., Smith W.W. Cold Spring Harbor Symp. Quant. Biol. 52:483-490(1987).

[2] Kumada Y., Benson D.R., Hillemann D., Hosted T.J., Rochefort D.A., Thompson C.J., Wohlleben W., Tateno Y. Proc. Natl. Acad. Sci. U.S.A. 90:3009-3013(1993).

[3] Shatters R.G., Kahn M.L. J. Mol. Evol. 29:422-428(1989).

[4] Brown J.R., Masuchi Y., Robb F.T., Doolittle W.F. J. Mol. Evol. 38:566-576(1994).

261. Globins profile (globin1)

Globins are heme-containing proteins involved in binding and/or transporting oxygen [1].

They belong to a very large and well studied family which is widely distributed in many organisms. The major groups of globins are: - Hemoglobins (Hb) from vertebrates. Hb is the protein responsible for transporting oxygen from the lungs to other tissues. It is a tetramer of two alpha and two beta chains. Most vertebrate species also express specific embryonic or fetal forms of hemoglobin where the alpha or the beta chains are replaced by a chain with higher oxygen affinity, as for the gamma, delta, epsilon and zeta chains in mammals, for

example. - Myoglobins (Mg) from vertebrates. Mg is a monomeric protein responsible for oxygen storage in muscles. - Invertebrate globins [2]. A wide variety of globins are found in invertebrates. Molluscs generally have one or two muscle globins which are either monomeric or dimeric. Insects, such as the midge *Chironomus thummi*, have a large set of extracellular globins. Nematodes and annelids have a variety of intracellular and extracellular globins; some of them are multi-domain polypeptides (from two up to nine-domain globins) and some produce large, disulfide-bonded aggregates. - Leghemoglobins (Lg) from the root nodules of leguminous plants. Lg provides oxygen for bacteroids. - Flavohemoproteins from bacteria (*Escherichia coli hmpA*) and fungi [3]. These proteins consist of two distinct domains: an N-terminal globin domain and a C-terminal FAD-containing reductase domain. In bacteria such as *Vitreoscilla*, the enzyme-associated globin is a single domain protein. All these globins seem to have evolved from a common ancestor. The profile developed to detect members of the globin family is based on a structural alignment of selected globin sequences [1] Concise Encyclopedia Biochemistry, Second Edition, Walter de Gruyter, Berlin New-York (1988).[2] Goodman M., Pedwaydon J., Czelusniak J., Suzuki T., Gotoh T., Moens L., Shishikura F., Walz D., Vinogradov S. J. Mol. Evol. 27:236-249(1988).

Plant hemoglobins signature (globin2)

Leghemoglobins [1] are hemoproteins present in the root nodules of leguminous plants. Leghemoglobins are structurally and functionally related to hemoglobin and myoglobin. By providing oxygen to the bacteroids, they are essential for symbiotic nitrogen fixation. Structurally related hemoglobins from the nodules of non-leguminous plants [2,3], and from the roots of non-nodulating plants[4] have been recently sequenced. A signature pattern was developed that picks up the sequence of plants hemoglobins, exclusively.

Consensus pattern: [SN]-P-x-L-x(2)-H-A-x(3)-F-

[1] Powell R., Gannon F. BioEssays 9:117-121(1988).

[2] Kortt A.A., Trinick M.J., Appleby C.A. Eur. J. Biochem. 175:141-149(1988).

[3] Kortt A.A., Inglis A.S., Fleming A.I., Appleby C.A. FEBS Lett. 231:341-346(1988).

[4] Bogusz D., Appleby C.A., Landsmann J., Dennis E.S., Trinick M.J., Peacock W.J. Nature 331:178-180(1988).

262. Fructose-bisphosphate aldolase class-I active site (glycolytic_enz)

Fructose-bisphosphate aldolase [1,2] is a glycolytic enzyme that catalyzes the reversible aldol cleavage or condensation of fructose-1,6-bisphosphate into dihydroxyacetone-phosphate and glyceraldehyde 3-phosphate. There are two classes of fructose-bisphosphate aldolases with different catalytic mechanisms. Class-I aldolases [3], mainly found in higher eukaryotes, are homotetrameric enzymes which form a Schiff-base intermediate between the C-2 carbonyl group of the substrate (dihydroxyacetone phosphate) and the epsilon-amino group of a lysine residue. In vertebrates, three forms of this enzyme are found: aldolase A in muscle, aldolase B in liver and aldolase C in brain. The sequence around the lysine involved in the Schiff-base is highly conserved and can be used as a signature for this class of enzyme.

Consensus pattern: [LIVM]-x-[LIVMFYW]-E-G-x-[LS]-L-K-P-[SN] [K is involved in Schiff-base formation]-

[1] Perham R.N. Biochem. Soc. Trans. 18:185-187(1990).

[2] Marsh J.J., Lebherz H.G. Trends Biochem. Sci. 17:110-113(1992).

[3] Freemont P.S., Dunbar B., Fothergill-Gilmore L.A. Biochem. J. 249:779-788(1988).

263. Glycosyl hydrolases family 11 active sites signatures

The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1,2]. Fungi and bacteria produces a spectrum of cellulolytic enzymes (cellulases) and xylanases which, on the basis of sequence similarities, can be classified into families. One of these families is known as the cellulase family G [3] or as the glycosyl hydrolases family 11 [4,E1]. The enzymes which are currently known to belong to this family are listed below. - *Aspergillus awamori* xylanase C (xynC). - *Bacillus circulans*, *pumilus*, *stearothermophilus* and *subtilis* xylanase (xynA). - *Clostridium acetobutylicum* xylanase (xynB). - *Clostridium stercorarium* xylanase A (xynA). - *Fibrobacter succinogenes* xylanase C (xynC) which consist of two catalytic domains that both belong to family 10. - *Neocallimastix patriciarum* xylanase A (xynA). - *Ruminococcus flavefaciens* bifunctional

xylanase XYLA (xynA). This protein consists of three domains: a N-terminal xylanase catalytic domain that belongs to family 11 of glycosyl hydrolases; a central domain composed of short repeats of Gln, Asn and Trp, and a C-terminal xylanase catalytic domain that belongs to family 10 of glycosyl hydrolases. - Schizophyllum commune xylanase A. - Streptomyces lividans xylanases B (xlnB) and C (xlnC). - Trichoderma reesei xylanases I and II. Two of the conserved regions in these enzymes are centered on glutamic acid residues which have both been shown [5], in Bacillus pumilis xylanase, to be necessary for catalytic activity. Both regions were used as signature patterns.

- 10 Consensus pattern: [PSA]-[LQ]-x-E-Y-Y-[LIVM](2)-[DE]-x-[FYWHN] [E is an active site residue]-
 Consensus pattern: [LIVMF]-x(2)-E-[AG]-[YWG]-[QRFGS]-[SG]-[STAN]-G-x-[SAF] [E is an active site residue]-

- 15 [1] Beguin P. Annu. Rev. Microbiol. 44:219-248(1990).
 [2] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).
 [3] Henrissat B., Claeyssens M., Tomme P., Lemesle L., Mornon J.-P. Gene 81:83-95(1989).
 [4] Henrissat B. Biochem. J. 280:309-316(1991).
 20 [5] Ko E.P., Akatsuka H., Moriyama H., Shinmyo A., Hata Y., Katsube Y., Urabe I., Okada H. Biochem. J. 288:117-121(1992).

264. Glycosyl hydrolase family 14

- 25 This family are beta amylases.

265. Glycosyl hydrolases family 1 signatures

- 30 It has been shown [1 to 4] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family: - Beta-glucosidases (EC 3.2.1.21) from various bacteria such as Agrobacterium strain ATCC 21400, Bacillus polymyxa, and Caldocellum saccharolyticum. - Two plants (clover) beta-glucosidases (EC 3.2.1.21). - Two different beta-galactosidases (EC 3.2.1.23) from the archaeobacteria Sulfolobus solfataricus

(genes bgaS and lacS). - 6-phospho-beta-galactosidases (EC 3.2.1.85) from various bacteria such as *Lactobacillus casei*, *Lactococcus lactis*, and *Staphylococcus aureus*. - 6-phospho-beta-glucosidases (EC 3.2.1.86) from *Escherichia coli* (genes bglB and ascB) and from *Erwinia chrysanthemi* (gene arbB). - Plants myrosinases (EC 3.2.3.1) (sinigrinase)

5 (thioglucosidase). - Mammalian lactase-phlorizin hydrolase (LPH) (EC 3.2.1.108 / EC 3.2.1.62). LPH, an integral membrane glycoprotein, is the enzyme that splits lactose in the small intestine. LPH is a large protein of about 1900 residues which contains four tandem repeats of a domain of about 450 residues which is evolutionary related to the above glycosyl hydrolases. One of the conserved regions in these enzymes is centered on a conserved
10 glutamic acid residue which has been shown [5], in the beta-glucosidase from *Agrobacterium*, to be directly involved in glycosidic bond cleavage by acting as a nucleophile. This region was used as a signature pattern. As a second signature pattern a conserved region was selected, found in the N-terminal extremity of these enzymes, this region also contains a glutamic acid residue.

Consensus pattern: [LIVMFSTC]-[LIVFYS]-[LIV]-[LIVMST]-E-N-G-[LIVMFAR]-
[CSAGN] [E is the active site residue]

Note: this pattern will pick up the last two domains of LPH; the first two domains, which are removed from the LPH precursor by proteolytic processing, have lost the active site
20 glutamate and may therefore be inactive [4].

Consensus pattern: F-x-[FYWM]-[GSTA]-x-[GSTA]-x-[GSTA](2)-[FYNH]-[NQ]-x-E-x-
[GSTA]-

[1] Henrissat B. Biochem. J. 280:309-316(1991).

25 [2] Henrissat B. Protein Seq. Data Anal. 4:61-62(1991).

[3] Gonzalez-Candelas L., Ramon D., Polaina J. Gene 95:31-38(1990).

[4] El Hassouni M., Henrissat B., Chippaux M., Barras F. J. Bacteriol. 174:765-777(1992).

[5] Withers S.G., Warren R.A.J., Street I.P., Rupitz K., Kempton J.B., Aebersold R. J. Am. Chem. Soc. 112:5887-5889(1990).

It has been shown [1,2,E1] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family: - Beta-galactosidases (EC 3.2.1.23) from bacteria such as *Escherichia coli* (genes *lacZ* and *ebgA*), *Clostridium acetobutylicum*, *Clostridium thermosulfurogenes*, *Klebsiella pneumoniae*, *Lactobacillus delbrueckii*, or

5 *Streptococcus thermophilus* and from the fungi *Kluyveromyces lactis*. - Beta-glucuronidase (EC 3.2.1.31) from *Escherichia coli* (gene *uidA*) and from mammals. One of the conserved regions in these enzymes is centered on a conserved glutamic acid residue which has been shown [3], in *Escherichia coli lacZ*, to be the general acid/base catalyst in the active site of the enzyme. This region was used as a signature pattern. As a second signature pattern a

10 highly conserved region was selected located some sixty residues upstream from the active site glutamate.

Consensus pattern: N-x-[LIVMFYWD]-R-[STACN](2)-H-Y-P-x(4)-[LIVMFYWS](2)-x(3)-[DN]-x(2)-G-[LIVMFYW](4)-

15 Consensus pattern: [DENQLF]-[KRVW]-N-[HRY]-[STAPV]-[SAC]-[LIVMFS](3)-W-[GS]-x(2,3)-N-E [E is the active site residue]-

[1] Henrissat B. *Biochem. J.* 280:309-316(1991).

[2] Schroeder C.J., Robert C., Lenzen G., McKay L.L., Mercenier A. J. *Gen. Microbiol.* 137:369-380(1991).

[3] Gebler J.C., Aebersold R., Withers S.G. *J. Biol. Chem.* 267:11126-11130(1992).

267. Glycosyl hydrolases family 3 active site

25 It has been shown [1,2] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:

- Beta glucosidases (EC 3.2.1.21) from the fungi *Aspergillus wentii* (A-3), *Hansenula anomala*, *Kluyveromyces fragilis*, *Saccharomycopsis fibuligera*, (BGL1 and BGL2), *Schizophyllum commune* and *Trichoderma reesei* (BGL1).
- 30 - Beta glucosidases from the bacteria *Agrobacterium tumefaciens* (Cbg1), *Butyrivibrio fibrisolvens* (*bglA*), *Clostridium thermocellum* (*bglB*), *Escherichia coli* (*bglX*), *Erwinia chrysanthemi* (*bgxA*) and *Ruminococcus albus*. - *Alteromonas* strain O-7 beta-hexosaminidase A (EC 3.2.1.52).

- *Bacillus subtilis* hypothetical protein yzbA.
- *Escherichia coli* hypothetical protein ycfO and HI0959, the corresponding *Haemophilus influenzae* protein.

One of the conserved regions in these enzymes is centered on a conserved aspartic acid residue which has been shown [3], in *Aspergillus wentii* beta-glucosidase A3, to be implicated in the catalytic mechanism. This region was used as a signature pattern.

Consensus pattern: [LIVM](2)-[KR]-x-[EQK]-x(4)-G-[LIVMFT]-[LIVT]-[LIVMF]-[ST]-D-x(2)-[SGADNI] [D is the active site residue]

[1] Henrissat B. Biochem. J. 280:309-316(1991).

[2] Castle L.A., Smith K.D., Morris R.O. J. Bacteriol. 174:1478-1486(1992).

[3] Bause E., Legler G. Biochim. Biophys. Acta 626:459-465(1980).

268. Glycosyl hydrolases family 8 signature

The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91)(exoglucanases), or xylanases (EC 3.2.1.8) [1,2]. Fungi and bacteria produces a spectrum of cellulolytic enzymes (cellulases) and xylanases which, on the basis of sequence similarities, can be classified into families. One of these families is known as the cellulase family D [3] or as the glycosyl hydrolases family 8 [4,E1]. The enzymes which are currently known to belong to this family are listed below. - *Acetobacter xylinum* endonuclease cmcAX. - *Bacillus* strain KSM-330 acidic endonuclease K (Endo-K). - *Cellulomonas josui* endoglucanase 2 (celB). - *Cellulomonas uda* endoglucanase. - *Clostridium cellulolyticum* endoglucanases C (celcCC). - *Clostridium thermocellum* endoglucanases A (celA). - *Erwinia chrysanthemi* minor endoglucanase y (celY). - *Bacillus circulans* beta-glucanase (EC 3.2.1.73). - *Escherichia coli* hypothetical protein yhjM. The most conserved region in these enzymes is a stretch of about 20 residues that contains two conserved aspartate. The first asparatate is thought [5] to act as the nucleophile in the catalytic mechanism. This region was used as a signature pattern.

Consensus pattern: A-[ST]-D-[AG]-D-x(2)-[IM]-A-x-[SA]-[LIVM]-[LIVMG]-x-A- x(3)-
[FW] [The first D is an active site residue]-

[1] Beguin P. Annu. Rev. Microbiol. 44:219-248(1990).

5 [2] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).

[3] Henrissat B., Claeysens M., Tomme P., Lemesle L., Mornon J.-P. Gene 81:83-95(1989).

[4] Henrissat B. Biochem. J. 280:309-316(1991).

[5] Alzari P.M., Souchon H., Dominguez R. Structure 4:265-275(1996).

10

269. Glycosyl hydrolases family 9 active sites signatures

The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1,2]. Fungi and bacteria produce a spectrum of cellulolytic enzymes (cellulases) and xylanases which, on the basis of sequence similarities, can be classified into families.

One of these families is known as the cellulase family E [3] or as the glycosyl hydrolases family 9 [4,E1]. The enzymes which are currently known to belong to this family are listed below. - *Butyrivibrio fibrisolvens* cellodextrinase 1 (ced1). - *Cellulomonas fimi*

20 endoglucanases B (cenB) and C (cenC). - *Clostridium cellulolyticum* endoglucanase G (celCCG). - *Clostridium cellulovorans* endoglucanase C (engC). - *Clostridium stercoararium* endoglucanase Z (avicelase I) (celZ). - *Clostridium thermocellum* endoglucanases D (celD), F (celF) and I (celI). - *Fibrobacter succinogenes* endoglucanase A (endA). - *Pseudomonas fluorescens* endoglucanase A (celA). - *Streptomyces reticuli* endoglucanase 1 (cel1). -

25 Thermomonospora fusca endoglucanase E-4 (celD). - *Dictyostelium discoideum* spore germination specific endoglucanase 270-6. This slime mold enzyme may digest the spore cell wall during germination, to release the enclosed amoeba. - Endoglucanases from plants such as Avocado or French bean. In plants this enzyme may be involved the fruit ripening process. Two of the most conserved regions in these enzymes are centered on conserved residues

30 which have been shown [5,6], in the endoglucanase D from *Cellulomonas thermocellum*, to be important for the catalytic activity. The first region contains an active site histidine and the second region contains two catalytically important residues: an aspartate and a glutamate.

Both regions were used as signature patterns.

Consensus pattern: [STV]-x-[LIVMFY]-[STV]-x(2)-G-x-[NKR]-x(4)-[PLIVM]-H-x-R [H is an active site residue]-

Consensus pattern: [FYW]-x-D-x(4)-[FYW]-x(3)-E-x-[STA]-x(3)-N-[STA] [D and E are active site residues]-

[1] Beguin P. Annu. Rev. Microbiol. 44:219-248(1990).

[2] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).

[3] Henrissat B., Claeysens M., Tomme P., Lemesle L., Mornon J.-P. Gene 81:83-95(1989).

[4] Henrissat B. Biochem. J. 280:309-316(1991).

[5] Tomme P., Chauvaux S., Beguin P., Millet J., Aubert J.-P., Claeysens M. J. Biol. Chem. 266:10313-10318(1991).

[6] Tomme P., van Beeumen J., Claeysens M. Biochem. J. 285:319-324(1992).

270. Glyceraldehyde 3-phosphate dehydrogenase active site (gpdh)

Glyceraldehyde 3-phosphate dehydrogenase (EC 1.2.1.12) (GAPDH) [1] is a tetrameric NAD-binding enzyme common to both the glycolytic and gluconeogenic pathways. A cysteine in the middle of the molecule is involved in forming a covalent phosphoglycerol thioester intermediate. The sequence around this cysteine is totally conserved in eubacterial and eukaryotic GAPDHs and is also present, albeit in a variant form, in the otherwise highly divergent archaeobacterial GAPDH [2]. Escherichia coli D-erythrose 4-phosphate dehydrogenase (E4PDH) (gene *epd* or *gapB*) is an enzyme highly related to GAPDH [3].

Consensus pattern: [ASV]-S-C-[NT]-T-x(2)-[LIM] [C is the active site residue]-

[1] Harris J.I., Waters M. (In) The Enzymes (3rd edition) 13:1-50(1976).

[2] Fabry S., Lang J., Niermann T., Vingron M., Hensel R. Eur. J. Biochem. 179:405-413(1989).

[3] Zhao G., Pease A.J., Bharani N., Winkler M.E. J. Bacteriol. 177:2804-2812(1995).

271. Granulins signature

Granulins [1] are a family of cysteine-rich peptides of about 6 Kd which may have multiple biological activity. A precursor protein (known as acrogranin) potentially encodes seven different forms of granulin (grnA to grnG) which are probably released by post-translational proteolytic processing. A schematic representation of the structure of a granulin is shown below: xxxCxxxxxCxxxxCCxxxxxxxxCCxxxxxCxxxxxCxxxxxCxxxxxCx
 *****'C': conserved cysteine probably involved in a disulfide bond.'*': position of the pattern. Granulins are evolutionary related to a PMP-D1, a peptide extracted from the pars intercerebralis of migratory locusts [2].

Consensus pattern: C-x-D-x(2)-H-C-C-P-x(4)-C [The four C's are probably involved in disulfide bonds]-

[1] Bhandari V., Palfree R.G., Bateman A. Proc. Natl. Acad. Sci. U.S.A. 89:1715-1719(1992).

[2] Nakakura N., Hietter H., van Dorsselaer A., Luu B. Eur. J. Biochem. 204:147-153(1992).

272. (HCV RdRp) Hepatitis C virus RNA dependent RNA polymerase

The RNA dependent RNA polymerase is also known as non-structural protein NS5B. NS5B is a 65 kDa protein that resembles other viral RNA polymerases. HCV replication is thought to occur in membrane bound replication complexes. These complexes transcribe the positive strand and the resulting minus strand is used as a template for the synthesis of genomic RNA. There are two viral proteins involved in the reaction, NS3 and NS5B.[1,2]

[1] Lohmann V, Korner F, Herian U, Bartenschlager R; J Virol 1997;71:8416-8428. [2] Behrens SE, Tomei L, De Francesco R; EMBO J 1996;15:12-22. [3] Ishido S, Fujita T, Hotta H; Biochem Biophys Res Commun 1998;244:35-40.

273. (HHH) Helix-hairpin-helix motif.

- 5 [1] Doherty AJ, Serpell LC, Ponting CP; Nucleic Acids Res 1996;24:2488-2497.

274. HIT family signature

- 10 Recently a family of small proteins of about 12 to 16 Kd has been described[1]. This family currently consists of: - Mammalian protein HINT (also known as Protein kinase C inhibitor 1 or PKCI- 1). HINT was incorrectly thought to be a specific inhibitor of PKC. It has been shown to bind zinc. - Fission yeast diadenosine 5',5'''-P1,P4-tetraphosphate asymmetrical hydrolase (Ap4Aase) (EC 3.6.1.17) [2] (gene aph1), which cleaves A-5'-PPPP- 5'A to yield AMP and ATP. - FHIT, a human protein whose gene is altered in different tumors and which acts [3] as a diadenosine 5',5'''-P1,P3-triphosphate hydrolase (Ap3Aase) (EC 3.6.1.29)
- 5 cleaving A-5'-PPP-5'A to yield AMP and ADP. - Yeast proteins HNT1 and HNT2. - Maize zinc-binding protein ZBP14. - Escherichia coli hypothetical protein ycfF. - Haemophilus influenzae hypothetical protein HI0961. - Helicobacter pylori hypothetical protein HP0404. - Methanococcus jannaschii hypothetical protein MJ0866. - Mycobacterium leprae
- 20 hypothetical protein U296A. - Synechocystis strain PCC 6803 hypothetical protein slr1234. - Caenorhabditis elegans hypothetical protein F21C3.3. - A hypothetical 13.2 Kd protein in hisE 3'region in Azospirillum brasilense. - A hypothetical 13.1 Kd protein in p37 5'region in Mycoplasma hyorhinitis. - A hypothetical 12.4 Kd protein in psbAII 5'region in Synechococcus strain PCC 7942. All these proteins contains a region with three clustered
- 25 histidines. This region is responsible for the designation of this family: HIT, for 'HistidineTriad' [1]. This region was originally thought to be implied in the binding of a zinc ion but was later identified [4] as part of the alpha-phosphate binding site of a nucleotide-binding domain. As a signature pattern, the region of the histidine triad was selected.
- 30 Consensus pattern: [NQA]-x(4)-[GAV]-x-[QF]-x-[LIVM]-x-H-[LIVMFYT]-H-[LIVMFT]-H-[LIVMF](2)-[PSGA]-

[1] Seraphin B. DNA Seq. 3:177-179(1992).

[2] Huang Y., Garrison P.N., Barnes L.D. *Biochem. J.* 312:925-932(1995).

[3] Barnes L.D., Garrison P.N., Siprashvili Z., Guranowski A., Robinson A.K., Ingram S.W., Croce C.M., Ohta M., Huebner K. *Biochemistry* 35:11529-11535(1996).

[4] Brenner C., Garrison P., Gilmour J., Peisach D., Ringe D., Petsko G.A., Lowenstein J.M.

5 Nat. Struct. Biol. 4:231-238(1997).

275. Myc-type, 'helix-loop-helix' dimerization domain signature (HLH)

A number of eukaryotic proteins, which probably are sequence specific DNA-binding
 10 proteins that act as transcription factors, share a conserved domain of 40 to 50 amino acid
 residues. It has been proposed [1] that this domain is formed of two amphipathic helices
 joined by a variable length linker region that could form a loop. This 'helix-loop-helix' (HLH)
 domain mediates protein dimerization and has been found in the proteins listed below
 [2,3,E1,E2]. Most of these proteins have an extra basic region of about 15 amino acid
 15 residues that is adjacent to the HLH domain and specifically binds to DNA. They are referred
 as basic helix-loop-helix proteins (bHLH), and are classified in two groups: class A
 (ubiquitous) and class B (tissue-specific). Members of the bHLH family bind variations on
 the core sequence 'CANNTG', also referred to as the E-box motif. The homo- or
 heterodimerization mediated by the HLH domain is independent of, but necessary for DNA
 20 binding, as two basic regions are required for DNA binding activity. The HLH proteins
 lacking the basic domain (Emc, Id) function as negative regulators since they form
 heterodimers, but fail to bind DNA. The hairy-related proteins (hairy, E(spl), deadpan) also
 repress transcription although they can bind DNA. The proteins of this subfamily act together
 with co-repressor proteins, like groucho, through their C-terminal motif WRPW. - The myc
 25 family of cellular oncogenes [4], which is currently known to contain four members: c-myc
 [E3], N-myc, L-myc, and B-myc. The myc genes are thought to play a role in cellular
 differentiation and proliferation. - Proteins involved in myogenesis (the induction of muscle
 cells). In mammals MyoD1 (Myf-3), myogenin (Myf-4), Myf-5, and Myf-6 (Mrf4 or
 herculin), in birds CMD1 (QMF-1), in *Xenopus* MyoD and MF25, in *Caenorhabditis elegans*
 30 CeMyoD, and in *Drosophila nautilus* (nau). - Vertebrate proteins that bind specific DNA
 sequences ('E boxes') in various immunoglobulin chains enhancers: E2A or ITF-1 (E12/pan-2
 and E47/pan-1), ITF-2 (tcf4), TFE3, and TFEB. - Vertebrate neurogenic differentiation factor
 1 that acts as differentiation factor during neurogenesis. - Vertebrate MAX protein, a

transcription regulator that forms a sequence- specific DNA-binding protein complex with myc or mad. - Vertebrate Max Interacting Protein 1 (MXI1 protein) which acts as a transcriptional repressor and may antagonize myc transcriptional activity by competing for max. - Proteins of the bHLH/PAS superfamily which are transcriptional activators. In

5 mammals, AH receptor nuclear translocator (ARNT), single-minded homologs (SIM1 and SIM2), hypoxia-inducible factor 1 alpha (HIF1A), AH receptor (AHR), neuronal pas domain proteins (NPAS1 and NPAS2), endothelial pas domain protein 1 (EPAS1), mouse ARNT2, and human BMAL1. In drosophila, single-minded (SIM), AH receptor nuclear translocator (ARNT), trachealess protein (TRH), and similar protein (SIMA). - Mammalian transcription

10 factors HES, which repress transcription by acting on two types of DNA sequences, the E box and the N box. - Mammalian MAD protein (max dimerizer) which acts as transcriptional repressor and may antagonize myc transcriptional activity by competing for max. - Mammalian Upstream Stimulatory Factor 1 and 2 (USF1 and USF2), which bind to a symmetrical DNA sequence that is found in a variety of viral and cellular promoters. -

15 Human lyl-1 protein; which is involved, by chromosomal translocation, in T- cell leukemia. - Human transcription factor AP-4. - Mouse helix-loop-helix proteins MATH-1 and MATH-2 which activate E box- dependent transcription in collaboration with E47. - Mammalian stem cell protein (SCL) (also known as tal1), a protein which may play an important role in hemopoietic differentiation. SCL is involved, by chromosomal translocation, in stem-cell leukemia. - Mammalian proteins Id1 to Id4 [5]. Id (inhibitor of DNA binding) proteins lack a

20 basic DNA-binding domain but are able to form heterodimers with other HLH proteins, thereby inhibiting binding to DNA. - Drosophila extra-macrochaetae (emc) protein, which participates in sensory organ patterning by antagonizing the neurogenic activity of the achaete- scute complex. Emc is the homolog of mammalian Id proteins. - Human Sterol

25 Regulatory Element Binding Protein 1 (SREBP-1), a transcriptional activator that binds to the sterol regulatory element 1 (SRE-1) found in the flanking region of the LDLR gene and in other genes. - Drosophila achaete-scute (AS-C) complex proteins T3 (l'sc), T4 (scute), T5 (achaete) and T8 (asense). The AS-C proteins are involved in the determination of the neuronal precursors in the peripheral nervous system and the central nervous system. -

30 Mammalian homologs of achaete-scute proteins, the MASH-1 and MASH-2 proteins. - Drosophila atonal protein (ato) which is involved in neurogenesis. - Drosophila daughterless (da) protein, which is essential for neurogenesis and sex-determination. - Drosophila deadpan (dpn), a hairy-like protein involved in the functional differentiation of neurons. - Drosophila

10

20

- 25

30

High mobility group (HMG) proteins are a family of relatively low molecular weight non-histone components in chromatin. HMG14 and HMG17 [1], two related proteins of about 100 amino acid residues, bind to the inner side of the nucleosomal DNA thus altering the

interaction between the DNA and the histone octamer. These two proteins may be involved in the process which maintains transcribable genes in a unique chromatin conformation. The trout nonhistone chromosomal protein H6 (histone T) also belongs to this family. As a signature pattern a conserved stretch of 10 residues located in the N-terminal section of HMG14 and HMG17 was selected.

Consensus pattern: R-R-S-A-R-L-S-A-[RK]-P-

[1] Bustin M., Reeves R. Prog. Nucleic Acid Res. Mol. Biol. 54:35-100(1996).

277. Hydroxymethylglutaryl-coenzyme A lyase active site (HMGL1)
3-hydroxy-3-methylglutaryl-coenzyme A lyase (HMG-CoA lyase or HL) (EC 4.1.3.4) catalyzes the transformation of HMG-CoA into acetyl-CoA and acetoacetate. In vertebrates it is a mitochondrial enzyme which is involved in ketogenesis and in leucine catabolism [1]. In some bacteria, such as *Pseudomonas mevalonii*, it is involved in mevalonate catabolism (gene *mvaB*). A cysteine has been shown [2], in *mvaB*, to be required for the activity of the enzyme. The region around this residue is perfectly conserved and is used as a signature pattern.

Consensus pattern: S-V-A-G-L-G-G-C-P-Y [C is the active site residue]-

[1] Mitchell G.A., Robert M.-F., Hruz P.W., Wang S., Fontaine G., Behnke C.E., Mende-Mueller L.M., Schappert K., Lee C., Gibson K.M., Miziorko H.M. J. Biol. Chem. 268:4376-4381(1993).

[2] Hruz P.W., Narasimhan C., Miziorko H.M. Biochemistry 31:6842-6847(1992).

Alpha-isopropylmalate and homocitrate synthases signatures (HMGL2)

The following enzymes have been shown [1] to be functionally as well as evolutionary related: - Alpha-isopropylmalate synthase (EC 4.1.3.12) which catalyzes the first step in the biosynthesis of leucine, the condensation of acetyl-CoA and alpha- ketoisovalerate to form 2-isopropylmalate synthase. - Homocitrate synthase (EC 4.1.3.21) (gene *nifV*) which is involved in the biosynthesis of the iron-molybdenum cofactor of nitrogenase and catalyzes

the condensation of acetyl-CoA and alpha-ketoglutarate into homocitrate. - Soybean late nodulin 56. - Methanococcus jannaschii hypothetical proteins MJ0503, MJ1195 and MJ1392. Two conserved regions were selected as signature patterns for these enzymes. The first region is located in the N-terminal section while the second region is located in the central section and contains two conserved histidine residues which could be implicated in the catalytic mechanism.

Consensus pattern: L-R-[DE]-G-x-Q-x(10)-K-

Consensus pattern: [LIVMFHW]-x(2)-H-x-H-[DN]-D-x-G-x-[GAS]-x-[GASLI]-

[1] Wang S.-Z., Dean D.R., Chen J.-S., Johnson J.L. J. Bacteriol. 173:3041-3046(1991).

278. (HMG CoA synt) Hydroxymethylglutaryl-coenzyme A synthase active site

Hydroxymethylglutaryl-coenzyme A synthase (EC 4.1.3.5) (HMG-CoA synthase) catalyzes the condensation of acetyl-CoA with acetoacetyl-CoA to produce HMG- CoA and CoA [1].In vertebrates there are two isozymes located in different subcellular compartments: a cytosolic form which is the starting point of the mevalonate pathway which leads to cholesterol and other sterolic and isoprenoid compounds and a mitochondrial form responsible for ketone body biosynthesis. HMG-CoA is also found in other eukaryotes such as insect, plants and fungi. A cysteine is known to act as the catalytic nucleophile in the first step of the reaction, the acetylation of the enzyme by acetyl-CoA. The conserved region was used around this active site residue as a signature pattern.

Consensus pattern: N-x-[DN]-[IV]-E-G-[IV]-D-x(2)-N-A-C-[FY]-x-G [C is the active site residue]-

[1] Rokosz L.L., Boulton D.A., Butkiewicz E.A., Sanyal G., Cueto M.A., Lachance P.A., Hermes J.D. Arch. Biochem. Biophys. 312:1-13(1994).

279. HMG (high mobility group) box

280. HSF-type DNA-binding domain signature

Heat shock factor (HSF) is a DNA-binding protein that specifically binds heat shock promoter elements (HSE). HSE is a palindromic element rich with repetitive purine and pyrimidine motifs: 5'-nGAAnnTTCnnGAAnnTTCn-3'. HSF is expressed at normal temperatures but is activated by heat shock or chemical stressors [1,2]. The sequences of HSF from various species show extensive similarity in a region of about 90 amino acids, which has been shown [3] to bind DNA. Some other proteins also contain a HSF domain, these are:

- Yeast SFL1, a protein involved in cell surface assembly and regulation of the gene related to flocculation (asexual cell aggregation) [4].
- Yeast transcription factor SKN7 (or BRY1 or POS9), which binds to the promoter elements SCB and MCB essential for the control of G1 cyclins expression [5].
- Yeast MGA1.
- Yeast hypothetical protein YJR147w.

A pattern from the most conserved part of the HSF DNA-binding domain was derived, its central region.

Consensus pattern: L-x(3)-[FY]-K-H-x-N-x-[STAN]-S-F-[LIVM]-R-Q-L-[NH]-x-Y-x-[FYW]-[RKH]-K-[LIVM]-

[1] Sorger P.K. Cell 65:363-366(1991).

[2] Mager W.H., Moradas Ferreira P. Biochem. J. 290:1-13(1993).

[3] Vuister G.W., Kim S.-J., Orosz A., Marquardt J., Wu C., Bax A. Nat. Struct. Biol. 1:605-613(1994).

[4] Fujita A., Kikuchi Y., Kuhara S., Misumi Y., Matsumoto S., Kobayashi H. Gene 85:321-328(1989).

[5] Morgan B.A., Bouquin N., Merrill G.F., Johnston L.H. EMBO J. 14:5679-5689(1995).

281. Heat shock hsp20 proteins family profile

Prokaryotic and eukaryotic organisms respond to heat shock or other environmental stress by inducing the synthesis of proteins collectively known as heat-shock proteins (hsp) [1].

Amongst them is a family of proteins with an average molecular weight of 20 Kd, known as the hsp20 proteins [2 to 5]. These seem to act as chaperones that can protect other proteins against heat-induced denaturation and aggregation. Hsp20 proteins seem to form large heterooligomeric aggregates; their family is currently composed of the following members: -

Vertebrate heat shock protein hsp27 (hsp25), induced by a variety of environmental stresses.

- *Drosophila* heat shock proteins hsp22, hsp23, hsp26, hsp27, hsp67BA and BC. -

Caenorhabditis elegans hsp16 multigene family. - Fungal HSP26 (budding yeast) and hsp30

(*Neurospora crassa* and *Aspergillus Nidulans*). - Plant small hsp's. Plants have four classes of

hsp20: classes I and II which are cytoplasmic, class III which is chloroplastic and class IV

which is found in the endomembrane. - Alpha-crystallin A and B chains. Alpha-crystallin is

an abundant constituent of the eye lens of most vertebrate species. Its main function appears

to be to maintain the correct refractive index of the lens. It is also found in other tissues

where it seems to act as a chaperone [6]. - *Schistosoma mansoni* major egg antigen p40.

Structurally, p40 is built of two tandem hsp20 domains. - A variety of prokaryotic proteins:

ibpA and ibpB from *Escherichia coli*, hsp18 from *Clostridium acetobutylicum*, spore protein

SP21 (hspA) from *Stigmatella aurantiaca*, *Mycobacterium leprae* 18 Kd antigen and

Mycobacterium tuberculosis 14 Kd antigen. - *Methanococcus jannaschii* hypothetical protein

MJ0285. Structurally, this family is characterized by the presence of a conserved C-terminal

domain of about 100 residues. The profile developed to detect members of the hsp20 family

is based on an alignment of this domain.

-Sequences known to belong to this class detected by the profile: ALL.

[1] Lindquist S., Craig E.A. *Annu. Rev. Genet.* 22:631-677(1988).[2] de Jong W.W.,

Leunissen J.A.M., Voorter C.E.M. *Mol. Biol. Evol.* 10:103-126(1993).[3] Caspers G.J.,

Leunissen J.A.M., de Jong W.W. *J. Mol. Evol.* 40:238-248(1995).[4] Jaenicke R., Creighton

T.E. *Curr. Biol.* 3:234-235(1993).[5] Jakob U., Buchner J. *Trends Biochem. Sci.* 19:205-

211(1994).[6] Groenen P.J.T.A., Merck K.B., de Jong W.W., Bloemendal H. *Eur. J.*

Biochem. 225:1-9(1994).

282. Heat shock hsp70 proteins family signatures

Prokaryotic and eukaryotic organisms respond to heat shock or other environmental stress by the induction of the synthesis of proteins collectively known as heat-shock proteins

(hsp) [1]. Amongst them is a family of proteins with an average molecular weight of 70 Kd,

known as the hsp70proteins [2,3,4]. In most species, there are many proteins that belong to

the hsp70 family. Some of them are expressed under unstressed conditions. Hsp70proteins

can be found in different cellular compartments (nuclear, cytosolic, mitochondrial,

endoplasmic reticulum, etc.). Some of the hsp70 family proteins are listed below: - In

Escherichia coli and other bacteria, the main hsp70 protein is known as the dnaK protein. A second protein, hscA, has been recently discovered. dnaK is also found in the chloroplast genome of red algae. - In yeast, at least ten hsp70 proteins are known to exist: SSA1 to SSA4, SSB1, SSB2, SSC1, SSD1 (KAR2), SSE1 (MSI3) and SSE2. - In Drosophila, there are at least eight different hsp70 proteins: HSP70, HSP68, and HSC-1 to HSC-6. - In mammals, there are at least eight different proteins: HSPA1 to HSPA6, HSC70, and GRP78 (also known as the immunoglobulin heavy chain binding protein (BiP)). - In the sugar beet yellow virus (SBYV), a hsp70 homolog has been shown [5] to exist. - In archaebacteria, hsp70 proteins are also present [6]. All proteins belonging to the hsp70 family bind ATP. A variety of functions has been postulated for hsp70 proteins. It now appears [7] that some hsp70 proteins play an important role in the transport of proteins across membranes. They also seem to be involved in protein folding and in the assembly/disassembly of protein complexes [8]. Three signature patterns for the hsp70 family of proteins were derived; the first centered on a conserved pentapeptide found in the N-terminal section of these proteins; the two others on conserved regions located in the central part of the sequence.

Consensus pattern: [IV]-D-L-G-T-[ST]-x-[SC] -

Consensus pattern: [LIVMF]-[LIVMFY]-[DN]-[LIVMFS]-G-[GSH]-[GS]-[AST]-x(3)-[ST]-[LIVM]-[LIVMFC]-

Consensus pattern: [LIVMY]-x-[LIVMF]-x-G-G-x-[ST]-x-[LIVM]-P-x-[LIVM]-x-[DEQKRSTA]-

[1] Lindquist S., Craig E.A. Annu. Rev. Genet. 22:631-677(1988).

[2] Pelham H.R.B. Cell 46:959-961(1986).

[3] Pelham H.R.B. Nature 332:776-77(1988).[4] Craig E.A. BioEssays 11:48-52(1989).

[5] Agranovsky A.A., Boyko V.P., Karasev A.V., Koonin E.V., Dolja V.V. J. Mol. Biol. 217:603-610(1991).

[6] Gupta R.S., Singh B. J. Bacteriol. 174:4594-4605(1992).

[7] Deshaies R.J., Koch B.D., Schekman R. Trends Biochem. Sci. 13:384-388(1988).

[8] Craig E.A., Gross C.A. Trends Biochem. Sci. 16:135-140(1991).

Prokaryotic and eukaryotic organisms respond to heat shock or other environmental stress by the induction of the synthesis of proteins collectively known as heat-shock proteins (hsp) [1]. Amongst them is a family of proteins, with an average molecular weight of 90 Kd, known as the hsp90proteins. Proteins known to belong to this family are: - Escherichia coli and other bacteria heat shock protein c62.5 (gene htpG). - Vertebrate hsp 90-alpha (hsp 86) and hsp 90-beta (hsp 84). - Drosophila hsp 82 (hsp 83). - Trypanosoma cruzi hsp 85. - Plants Hsp82 or Hsp83. - Yeast and other fungi HSC82, and HSP82. - The endoplasmic reticulum protein 'endoplasmin' (also known as Erp99 in mouse, GRP94 in hamster, and hsp 108 in chicken). The exact function of hsp90 proteins is not yet known. In higher eukaryotes, hsp90 has been found associated with steroid hormone receptors, with tyrosine kinase oncogene products of several retroviruses, with eIF2alpha kinase, and with actin and tubulin. Hsp90 are probable chaperonins that possess ATPase activity [2,3]. As a signature pattern for the hsp90 family of proteins, a highly conserved region found in the N-terminal part of these proteins was selected.

Consensus pattern: Y-x-[NQH]-K-[DE]-[IVA]-F-L-R-[ED] -

[1] Lindquist S., Craig E.A. Annu. Rev. Genet. 22:631-677(1988).

[2] Nadeau K., Das A., Walsh C.T. J. Biol. Chem. 268:1479-1487(1993).

[3] Jakob U., Buchner J. Trends Biochem. Sci. 19:205-211(1994).

284. Helix-turn-helix (HTH3)

This large family of DNA binding helix-turn helix proteins includes Cro

Swiss:P03036 and CI Swiss:P03034.

285. Heme oxygenase signature

Heme oxygenase (EC 1.14.99.3) (HO) [1] is the microsomal enzyme that, in animals, carries out the oxidation of heme, it cleaves the heme ring at the alpha methene bridge to form biliverdin and carbon monoxide. Biliverdin is subsequently converted to bilirubin by biliverdin reductase. In mammals there are three isozymes of heme oxygenase: HO-1 to HO-3. The first two isozymes differ in their tissue expression and their inducibility: HO-1 is

highly inducible by its substrate heme and by various non-heme substances, while HO-2 is non-inducible. It has been suggested [2] that HO-2 could be implicated in the production of carbon monoxide in the brain where it is said to act as a neurotransmitter. In the genome of the chloroplast of red algae as well as in cyanobacteria, there is a heme oxygenase (gene pbsA) that is the key enzyme in the synthesis of the chromophoric part of the photosynthetic antennae [3]. An heme oxygenase is also present in the bacteria *Corynebacterium diphtheriae* (gene hmuO), where it is involved in the acquisition of iron from the host heme [4]. There is, in the central section of these enzymes, a well conserved region centered on a histidine residue which is proposed to play a key role in binding the substrate heme at the active center of the enzyme. This region was used as a signature pattern.

Consensus pattern: L-[IV]-A-H-[STACH]-Y-[STV]-[RT]-Y-[LIVM]-G [H binds the heme] -

[1] Maines M.D. FASEB J. 2:2557-2568(1988).

[2] Barinaga M. Science 259:309-309(1993).

[3] Richaud C., Zabulon G. Proc. Natl. Acad. Sci. U.S.A. 94:11736-11741(1997).

[4] Schmitt M.P. J. Bacteriol. 179:838-845(1997).

286. Hepatitis core antigen.

The core antigen of hepatitis viruses possesses a carboxyl terminus rich in arginine. On this basis it was predicted that the core antigen would bind DNA [1]. There is some experimental evidence to support this [2].

[1] Pasek M, Goto T, Gilbert W, Zink B, Schaller H, McKay P, Leadbetter G, Murray K; Nature 1979;282:575-579. [2] Gallina A, Bonelli F, Zentilin L, Rindi G, Muttini M, Milanesi G; J Virol 1989;63:4645-4652.

287. Histidine biosynthesis protein

Proteins involved in steps 4 and 6 of the histidine biosynthesis pathway are contained in this family. Histidine is formed by several complex and distinct biochemical reactions catalysed by eight enzymes. The enzymes in this Pfam entry are called His6 and His7 in eukaryotes and HisA and HisF in prokaryotes.

- 5 [1] Fani R, Tamburini E, Mori E, Lazcano A, Lio P, Barberio C, Casalone E, Cavalieri D, Perito B, Polsinelli M, Gene 1997;197:9-17. [2] Fani R, Lio P, Chiarelli I, Bazzicalupo M, J Mol Evol 1994;38:489-495.

10 288. Histone deacetylase family

Histones can be reversibly acetylated on several lysine residues. Regulation of transcription is caused in part by this mechanism. Histone deacetylases catalyse the removal of the acetyl group. Histone deacetylases are related to other proteins [1].

Leipe DD, Landsman D, Nucleic Acids Res 1997;25:3693-3697.

15 289. Histidinol dehydrogenase signature

Histidinol dehydrogenase (EC 1.1.1.23) (HDH) catalyzes the terminal step in the biosynthesis of histidine in bacteria, fungi, and plants, the four-electron oxidation of L-histidinol to histidine. In bacteria HDH is a single chain polypeptide; in fungi it is the C-terminal domain of a multifunctional enzyme which catalyzes three different steps of histidine biosynthesis; and in plants it is expressed as nuclear encoded protein precursor which is exported to the chloroplast [1]. As a signature pattern a highly conserved region located in the central part of HDH was selected. This region does not correspond to the part of the enzyme that, in most, but not all HDH sequences contains a cysteine residue which, in *Salmonella typhimurium*, has been said [2] to be important for the catalytic activity of the enzyme.

Consensus pattern: I-D-x(2)-A-G-P-[ST]-E-[LIVS]-[LIVMA](3)-[AC]-x(3)-A-x(4)-[LIVM]-[AV]-[SACL]-[DE]-[LIVMFC]-[LIVM]-[SA]-x(2)-E-H-

- 30 [1] Nagai A., Ward E., Beck J., Tada S., Chang J.-Y., Scheidegger A., Ryals J. Proc. Natl. Acad. Sci. U.S.A. 88:4133-4137(1991).
[2] Grubmeyer C.T., Gray W.R. Biochemistry 25:4778-4784(1986).

290. Homoserine dehydrogenase signature

Homoserine dehydrogenase (EC 1.1.1.3) (HDh) [1,2] catalyzes NAD-dependent reduction of aspartate beta-semialdehyde into homoserine. This reaction is the third step in a pathway leading from aspartate to homoserine. The latter participates in the biosynthesis of threonine and then isoleucine as well as in that of methionine. HDh is found either as a single chain protein as in some bacteria and yeast, or as a bifunctional enzyme consisting of an N-terminal aspartokinase domain and a C-terminal HDh domain as in bacteria such as *Escherichia coli* and in plants. As a signature pattern, the best conserved region of Hdh has been selected. This is a segment of 23 to 24 residues located in the central section and that contains two conserved aspartate residues.

Consensus pattern: A-x(3)-G-[LIVMFY]-[STAG]-x(2,3)-[DNS]-P-x(2)-D-[LIVM]-x-G- x-D-x(3)-K-

[1] Thomas D., Barbey R., Surdin-Kerjan Y. FEBS Lett. 323:289-293(1993).

[2] Cami B., Clepet C., Patte J.-C. Biochimie 75:487-495(1993).

291. haloacid dehalogenase-like hydrolase

This family is structurally different from the alpha/ beta hydrolase family (abhydrolase). This family includes L-2-haloacid dehalogenase, epoxide hydrolases and phosphatases. The structure of the family consists of two domains. One is an inserted four helix bundle, which is the least well conserved region of the alignment, between residues 16 and 96 of Swiss:P24069. The rest of the fold is composed of the core alpha/beta domain.

[1] Hisano T, Hata Y, Fujii T, Liu JQ, Kurihara T, Esaki N, Soda K, J Biol Chem 1996; 271:20322-20330.

292. DEAD and DEAH box families ATP-dependent helicases signatures (helicase_C)

A number of eukaryotic and prokaryotic proteins have been characterized [1,2,3] on the basis of their structural similarity. They all seem to be involved in ATP-dependent, nucleic-acid

unwinding. Proteins currently known to belong to this family are: - Initiation factor eIF-4A. Found in eukaryotes, this protein is a subunit of a high molecular weight complex involved in 5'cap recognition and the binding of mRNA to ribosomes. It is an ATP-dependent RNA-helicase. - PRP5 and PRP28. These yeast proteins are involved in various ATP-requiring

5 steps of the pre-mRNA splicing process. - P110, a mouse protein expressed specifically during spermatogenesis. - An3, a *Xenopus* putative RNA helicase, closely related to P110. - SPP81/DED1 and DBP1, two yeast proteins probably involved in pre-mRNA splicing and related to P110. - *Caenorhabditis elegans* helicase glh-1. - MSS116, a yeast protein required for mitochondrial splicing. - SPB4, a yeast protein involved in the maturation of 25S

10 ribosomal RNA. - p68, a human nuclear antigen. p68 has ATPase and DNA-helicase activities in vitro. It is involved in cell growth and division. - Rm62 (p62), a *Drosophila* putative RNA helicase related to p68. - DBP2, a yeast protein related to p68. - DHH1, a yeast protein. - DRS1, a yeast protein involved in ribosome assembly. - MAK5, a yeast protein involved in maintenance of dsRNA killer plasmid. - ROK1, a yeast protein. - ste13, a fission yeast protein. - Vasa, a *Drosophila* protein important for oocyte formation and specification of embryonic posterior structures. - Me31B, a *Drosophila* maternally expressed protein of unknown function. - dbpA, an *Escherichia coli* putative RNA helicase. - deaD, an *Escherichia coli* putative RNA helicase which can suppress a mutation in the rpsB gene for ribosomal protein S2. - rhIB, an *Escherichia coli* putative RNA helicase. - rhIE, an *Escherichia coli* putative RNA helicase. - srmB, an *Escherichia coli* protein that shows RNA-dependent ATPase activity. It probably interacts with 23S ribosomal RNA. - *Caenorhabditis elegans* hypothetical proteins T26G10.1, ZK512.2 and ZK686.2. - Yeast hypothetical protein YHR065c. - Yeast hypothetical protein YHR169w. - Fission yeast hypothetical protein SpAC31A2.07c. - *Bacillus subtilis* hypothetical protein yxiN. All these proteins share a

25 number of conserved sequence motifs. Some of them are specific to this family while others are shared by other ATP-binding proteins or by proteins belonging to the helicases 'superfamily' [4,E1]. One of these motifs, called the 'D-E-A-D-box', represents a special version of the B motif of ATP-binding proteins. Some other proteins belong to a subfamily which have His instead of the second Asp and are thus said to be 'D-E-A-H-box' proteins

30 [3,5,6,E1]. Proteins currently known to belong to this subfamily are: - PRP2, PRP16, PRP22 and PRP43. These yeast proteins are all involved in various ATP-requiring steps of the pre-mRNA splicing process. - Fission yeast prh1, which may be involved in pre-mRNA splicing. - Male-less (mle), a *Drosophila* protein required in males, for dosage compensation of X

chromosome linked genes. - RAD3 from yeast. RAD3 is a DNA helicase involved in excision repair of DNA damaged by UV light, bulky adducts or cross-linking agents. Fission yeast rad15 (rhp3) and mammalian DNA excision repair protein XPD (ERCC-2) are the homologs of RAD3. - Yeast CHL1 (or CTF1), which is important for chromosome transmission and normal cell cycle progression in G(2)/M. - Yeast TPS1. - Yeast hypothetical protein YKL078w. - Caenorhabditis elegans hypothetical proteins C06E1.10 and K03H1.2. - Poxviruses' early transcription factor 70 Kd subunit which acts with RNA polymerase to initiate transcription from early gene promoters. - I8, a putative vaccinia virus helicase. - hrpA, an Escherichia coli putative RNA helicase. Signature patterns were developed for both subfamilies.

Consensus pattern: [LIVMF](2)-D-E-A-D-[RKEN]-x-[LIVMFYGSTN]-

Consensus pattern: [GSAH]-x-[LIVMF](3)-D-E-[ALIV]-H-[NECR] -

Note: proteins belonging to this family also contain a copy of the ATP/GTP- binding motif 'A' (P-loop) (see the relevant entry <[PDOC00017](#)

[1] Schmid S.R., Linder P. Mol. Microbiol. 6:283-292(1992).

[2] Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K., Schnier J., Slonimski P.P. Nature 337:121-122(1989).

[3] Wassarman D.A., Steitz J.A. Nature 349:463-464(1991).

[4] Hodgman T.C. Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).

[5] Harosh I., Deschavanne P. Nucleic Acids Res. 19:6331-6331(1991).

[6] Koonin E.V., Senkevich T.G. J. Gen. Virol. 73:989-993(1992).

293. Heme-binding domain in cytochrome b5 and oxidoreductases (heme_1)

Cytochrome b5 is a membrane-bound hemo protein which acts as an electron carrier for several membrane-bound oxygenases [1]. There are two homologous forms of b5, one found in microsomes and one found in the outer membrane of mitochondria. Two conserved histidine residues serve as axial ligands for the heme group. The structure of a number of oxidoreductases consists of the juxtaposition of a heme-binding domain homologous to that of b5 and either a flavodehydrogenase or a molybdopterin domain. These enzymes are:

- Lactate dehydrogenase (EC 1.1.2.3) [2], an enzyme that consists of a flavodehydrogenase domain and a heme-binding domain called cytochrome b2.
- Nitrate reductase (EC 1.6.6.1), a key enzyme involved in the first step of nitrate assimilation in plants, fungi and bacteria [3,4]. Consists of a molybdopterin domain (see <PDOC00484>), a heme-binding domain called cytochrome b557, as well as a cytochrome reductase domain.
- Sulfite oxidase (EC 1.8.3.1) [5], which catalyzes the terminal reaction in the oxidative degradation of sulfur-containing amino acids. Also consists of a molybdopterin domain and a heme-binding domain.

This family of proteins also includes:

- TU-36B, a Drosophila muscle protein of unknown function [6].
- Fission yeast hypothetical protein SpAC1F12.10c.
- Yeast hypothetical protein YMR073c.
- Yeast hypothetical protein YMR272c.

A segment was used which includes the first of the two histidine heme ligands, as a signature pattern for the heme-binding domain of cytochrome b5 family.

Consensus pattern: [FY]-[LIVMK]-x(2)-H-P-[GA]-G [H is a heme axial ligand]-

[1] Ozols J. Biochim. Biophys. Acta 997:121-130(1989).

[2] Guiard B. EMBO J. 4:3265-3272(1985).

[3] Calza R., Huttner E., Vincentz M., Rouze P., Galangau F., Vaucheret H., Cherel I., Meyer C., Kronenberger J., Caboche M. Mol. Gen. Genet. 209:552-562(1987).

[4] Crawford N.M., Smith M., Bellissimo D., Davis R.W. Proc. Natl. Acad. Sci. U.S.A.

85:5006-5010(1988).

[5] Guiard B., Lederer F. Eur. J. Biochem. 100:441-453(1979).

[6] Levin R.J., Boychuk P.L., Croniger C.M., Kazzaz J.A., Rozek C.E. Nucleic Acids Res. 17:6349-6367(1989).

294. Hexapeptide-repeat containing-transferases signature

On the basis of sequence similarity, a number of transferases have been proposed [1,2,3,4] to belong to a single family. These proteins are: - Serine acetyltransferase (EC 2.3.1.30) (SAT)

(gene *cysE*), an enzyme involved in cysteine biosynthesis. - *Azotobacter chroococcum* nitrogen fixation protein *nifP*. *NifP* is most probably a SAT involved in the optimization of nitrogenase activity. - *Escherichia coli* thiogalactoside acetyltransferase (EC 2.3.1.18) (gene *lacA*), an enzyme involved in the biosynthesis of lactose. - UDP-N-acetylglucosamine acyltransferase (EC 2.3.1.129) (gene *lpxA*), an enzyme involved in the biosynthesis of lipid A, a phosphorylated glycolipid that anchors the lipopolysaccharide to the outer membrane of the cell. - UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.-) (gene *lpxD* or *firA*), which is also involved in the biosynthesis of lipid A. - Chloramphenicol acetyltransferase (CAT) (EC 2.3.1.28) from *Agrobacterium tumefaciens*, *Bacillus sphaericus*, *Escherichia coli* plasmid IncFII NR79, *Pseudomonas aeruginosa*, *Staphylococcus aureus* plasmid pIP630. These CAT are not evolutionary related to the main family of CAT (see <PDOC00093>). - *Rhizobium* nodulation protein *nodL*. *NodL* is an acetyltransferase involved in the O-acetylation of Nod factors. - Bacterial maltose O-acetyltransferase (EC 2.3.1.79). - Bacterial tetrahydrodipicolinate N-succinyltransferase (EC 2.3.1.117) (gene *dapD*) which catalyzes the fourth step in the biosynthesis of diaminopimelate and lysine from aspartate semialdehyde. - Bacterial N-acetylglucosamine-1-phosphate uridyltransferase (EC 2.7.7.23) (gene *glmU* or *gcaD* or *tms*), an enzyme involved in peptidoglycan and lipopolysaccharide biosynthesis. - *Staphylococcus aureus* protein *capG* which is involved in biosynthesis of type 1 capsular polysaccharide. - Yeast hypothetical protein YJL218w, which is highly similar to *Escherichia coli* *lacA*. - Fission yeast hypothetical protein SpAC18B11.09c. - *Methanococcus jannaschii* hypothetical protein MJ1064. These proteins have been shown [3,4] to contain a repeat structure composed of tandem repeats of a [LIV]-G-x(4) hexapeptide which, in the tertiary structure of *lpxA* [5], has been shown to form a left-handed parallel beta helix. Our signature pattern is based on a fourfold repeat of this hexapeptide.

Consensus pattern: [LIV]-[GAED]-x(2)-[STAV]-x-[LIV]-x(3)-[LIVAC]-x-[LIV]-[GAED]-x(2)-[STAVR]-x-[LIV]-[GAED]-x(2)-[STAV]-x-[LIV]-x(3)-[LIV]-

[1] Downie J.A. Mol. Microbiol. 3:1649-1651(1989).

[2] Parent R., Roy P.H. J. Bacteriol. 174:2891-2897(1992).

[3] Vaara M. FEMS Microbiol. Lett. 97:249-254(1992).

[4] Vuorio R., Haerkonen T., Tolvanen M., Vaara M. FEBS Lett. 337:289-292(1994).

[5] Raetz C.R.H., Roderick S.L. Science 270:997-1000(1995).

295. Hexokinases signature. Hexokinase (EC 2.7.1.1) [1,2] is an important glycolytic enzyme that catalyzes the phosphorylation of keto- and aldohexoses (e.g. glucose, mannose and fructose) using MgATP as the phosphoryl donor. In vertebrates there are four major isoenzymes, commonly referred as types I,II, III and IV. Type IV hexokinase, which is often incorrectly designated glucokinase [3], is only expressed in liver and pancreatic beta-cells and plays an important role in modulating insulin secretion; it is a protein of a molecular mass of about 50 Kd. Hexokinases of types I to III, which have low Km values for glucose, have a molecular mass of about 100 Kd. Structurally they consist of a very small N-terminal hydrophobic membrane-binding domain followed by two highly similar domains of 450 residues. The first domain has lost its catalytic activity and has evolved into a regulatory domain. In yeast there are three different isozymes: hexokinase PI (gene HXK1), PII(gene HXKB), and glucokinase (gene GLK1). All three proteins have a molecular mass of about 50 Kd. All these enzymes contain one (or two in the case of types I to III isozymes)strongly conserved region which has been shown [4] to be involved in substrate binding. A pattern from that region has been derived

Consensus pattern: [LIVM]-G-F-[TN]-F-S-[FY]-P-x(5)-[LIVM]-[DNST]-x(3)-[LIVM]- x(2)-W-T-K-x-[LF]-

[1] Middleton R.J. Biochem. Soc. Trans. 18:180-183(1990).[2] Griffin L.D., Gelb B.D., Wheeler D.A., Davison D., Adams V., McCabe E.R. Genomics 11:1014-1024(1991).[3] Cornish-Bowden A., Luz Cardenas M. Trends Biochem. Sci. 16:281-282(1991).[4] Schirch D.M., Wilson J.E. Arch. Biochem. Biophys. 254:385-396(1987).

296. Histone H2A signature (his1)

Histone H2A is one of the four histones, along with H2B, H3 and H4, which forms the eukaryotic nucleosome core. Using alignments of histone H2Asequences [1,2,E1] as a signature pattern, a conserved region in the N-terminal part of H2A. This region is conserved

both in classical S-phase regulated H2A's and in variant histone H2A's which are synthesized throughout the cell cycle.

Consensus pattern: [AC]-G-L-x-F-P-V-

5

[1] Wells D.E., Brown D. Nucleic Acids Res. 19:2173-2188(1991).

[2] Thatcher T.H., Gorovsky M.A. Nucleic Acids Res. 22:174-179(1994).

Histone H4 signature (his2)

10

Histone H4 is one of the four histones, along with H2A, H2B and H3, which forms the eukaryotic nucleosome core. Along with H3, it plays a central role in nucleosome formation. The sequence of histone H4 has remained almost invariant in more than 2 billion years of evolution [1,E1]. The region used as a signature pattern is a pentapeptide found in positions 14 to 18 of all H4sequences. It contains a lysine residue which is often acetylated [2] and a histidine residue which is implicated in DNA-binding [3].

5

Consensus pattern: G-A-K-R-H-

[1] Thatcher T.H., Gorovsky M.A. Nucleic Acids Res. 22:174-179(1994).

[2] Doenecke D., Gallwitz D. Mol. Cell. Biochem. 44:113-128(1982).

[3] Ebralidse K.K., Grachev S.A., Mirzabekov A.D. Nature 331:365-367(1988).

20

Histone H3 signatures (his3)

Histone H3 is one of the four histones, along with H2A, H2B and H4, which forms the eukaryotic nucleosome core. It is a highly conserved protein of 135 amino acid residues [1,2,E1]. The following proteins have been found to contain a C-terminal H3-like domain: - Mammalian centromeric protein CENP-A [3]. Could act as a core histone necessary for the assembly of centromeres. - Yeast chromatin-associated protein CSE4 [4]. - Caenorhabditis elegans chromosome III encodes two highly related proteins (F54C8.2 and F58A4.3) whose C-terminal section is evolutionary related to the last 100 residues of H3. The function of these proteins is not yet known. Two signature patterns were developed, The first one corresponds to a perfectly conserved heptapeptide in the N-terminal part of H3. The second one is derived from a conserved region in the central section of H3.

25

30

Consensus pattern: K-A-P-R-K-Q-L-

Consensus pattern: P-F-x-[RA]-L-[VA]-[KRQ]-[DEG]-[IV]-

- 5 [1] Wells D.E., Brown D. Nucleic Acids Res. 19:2173-2188(1991).
- [2] Thatcher T.H., Gorovsky M.A. Nucleic Acids Res. 22:174-179(1994).
- [3] Sullivan K.F., Hechenberger M., Masri K. J. Cell Biol. 127:581-592(1994).
- [4] Stoler S., Keith K.C., Curnick K.E., Fitzgerald-Hayes M. Genes Dev. 9:573-586(1995).

10 Histone H2B signature (his4)

Histone H2B is one of the four histones, along with H2A, H3 and H4, which forms the eukaryotic nucleosome core. Using alignments of histone H2B sequences [1,2,E1], a conserved region was selected in the C-terminal part of H2B.

15 Consensus pattern: [KR]-E-[LIVM]-[EQ]-T-x(2)-[KR]-x-[LIVM](2)-x-[PAG]-[DE]-L- x-[KR]-H-A-[LIVM]-[STA]-E-G-

- [1] Wells D.E., Brown D. Nucleic Acids Res. 19:2173-2188(1991).
- [2] Thatcher T.H., Gorovsky M.A. Nucleic Acids Res. 22:174-179(1994).

20

297. 'Homeobox' domain signature and profile (home1)

The 'homeobox' is a protein domain of 60 amino acids [1 to 5,E1] first identified in a number of *Drosophila* homeotic and segmentation proteins. It has since been found to be extremely well conserved in many other animals, including vertebrates. This domain binds DNA through a helix-turn-helix type of structure. Some of the proteins which contain a homeobox domain play an important role in development. Most of these proteins are known to be sequence specific DNA-binding transcription factors. The homeobox domain has also been found to be very similar to a region of the yeast mating type proteins. These are sequence-specific DNA-binding proteins that act as master switches in yeast differentiation by controlling gene expression in a cell type-specific fashion. A schematic representation of the homeobox domain is shown below. The helix-turn-helix region is shown by the symbols 'H' (for helix), and 't' (for turn).

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxHHHHHHHHHtttHHHHHHHHHxxxxxxxx ||||| 1

10 20 30 40 50 60 The pattern to detect homeobox sequences that was developed is 24 residues long and spans positions 34 to 57 of the homeobox domain.

5 Consensus pattern: [LIVMFYG]-[ASLVR]-x(2)-[LIVMSTACN]-x-[LIVM]-x(4)-[LIV]-[RKNQESTAIY]-[LIVFSTNKH]-W-[FYVC]-x-[NDQTAH]-x(5)- [RKNAIMW] -

[1] Gehring W.J. (In) Guidebook to the homeobox genes, Duboule D., Ed., pp1-10, Oxford University Press, Oxford, (1994).

10 [2] Buerklin T.R. (In) Guidebook to the homeobox genes, Duboule D., Ed., pp25-72, Oxford University Press, Oxford, (1994).

[3] Gehring W.J. Trends Biochem. Sci. 17:277-280(1992).

[4] Gehring W.J., Hiromi Y. Annu. Rev. Genet. 20:147-173(1986).

[5] Schofield P.N. Trends Neurosci. 10:3-6(1987).

15 'Homeobox' antennapedia-type protein signature (home2)

The homeotic Hox proteins are sequence-specific transcription factors. They are part of a developmental regulatory system that provides cells with specific positional identities on the anterior-posterior (A-P) axis [1]. The hox proteins contain a 'homeobox' domain. In

20 Drosophila and other insects, there are eight different Hox genes that are encoded in two gene complexes, ANT-C and BX-C. In vertebrates there are 38 genes organized in four complexes.

In six of the eight Drosophila Hox genes the homeobox domain is highly similar and a conserved hexapeptide is found five to sixteen amino acids upstream of the homeobox

25 domain. The six Drosophila proteins that belong to this group are antennapedia (Antp), abdominal-A (abd-A), deformed (Dfd), proboscipedia (pb), sex combs reduced (scr) and ultrabithorax (ubx) and are collectively known as the 'antennapedia' subfamily. In vertebrates

the corresponding Hox genes are known [2] as Hox-A2, A3, A4, A5, A6, A7, Hox-B1, B2,

B3, B4, B5, B6, B7, B8, Hox-C4, C5, C6, C8, Hox-D1, D3, D4 and D8. *Caenorhabditis*

elegans lin-39 and mab-5 are also members of the 'antennapedia' subfamily. As a signature

30 pattern for this subfamily of homeobox proteins, the conserved hexapeptide was used.

Consensus pattern: [LIVMFE]-[FY]-P-W-M-[KRQTA]-

[1] McGinnis W., Krumlauf R. Cell 68:283-302(1992).

[2] Scott M.P. Cell 71:551-553(1992).

'Homeobox' engrailed-type protein signature (home3)

Most proteins which contain a 'homeobox' domain can be classified [1,2], on the basis of their sequence characteristics, in three subfamilies: engrailed, antenapedia and paired.

Proteins currently known to belong to the engrailed subfamily are: - Drosophila segmentation polarity protein engrailed (en) which specifies the body segmentation pattern and is required for the development of the central nervous system. - Drosophila invected protein (inv). - Silk

moth proteins engrailed and invected, which may be involved in the compartmentalization of the silk gland. - Honeybee E30 and E60. - Grasshopper (*Schistocerca americana*) G-En. -

Mammalian and birds En-1 and En-2. - Zebrafish Eng-1, -2 and -3. - Sea urchin (*Tripneustes gratilla*) SU-HB-en. - Leech (*Helobdella triserialis*) Ht-En. - *Caenorhabditis elegans* ceh-

16. Engrailed homeobox proteins are characterized by the presence of a conserved region of some 20 amino-acid residues located at the C-terminal of the 'homeobox' domain. As a signature pattern for this subfamily of proteins, a stretch of eight perfectly conserved residues in this region was used.

Consensus pattern: L-M-A-[EQ]-G-L-Y-N-

[1] Scott M.P., Tamkun J.W., Hartzell G.W. III *Biochim. Biophys. Acta* 989:25-48(1989).

[2] Gehring W.J. *Science* 236:1245-1252(1987).

298. Isocitrate lyase signature (ICL)

Isocitrate lyase (EC 4.1.3.1) [1,2] is an enzyme that catalyzes the conversion of isocitrate to succinate and glyoxylate. This is the first step in the glyoxylate bypass, an alternative to the tricarboxylic acid cycle in bacteria, fungi and plants. A cysteine, a histidine and a glutamate or aspartate have been found to be important for the enzyme's catalytic activity. Only one cysteine residue is conserved between the sequences of the fungal, plant and bacterial enzymes; it is located in the middle of a conserved hexapeptide that can be used as a signature pattern for this type of enzyme.

Consensus pattern: K-[KR]-C-G-H-[LMQ] [C is a putative active site residue]-

[1] Beeching J.R. Protein Seq. Data Anal. 2:463-466(1989).

[2] Atomi H., Ueda M., Hikida M., Hishida T., Teranishi Y., Tanaka A. J. Biochem.

5 107:262-266(1990).

299. Initiation factor 2 subunit

10 This family includes initiation factor 2B alpha, beta and delta subunits from eukaryotes, related proteins from archaeobacteria and IF-2 from prokaryotes. Initiation factor 2 binds to Met-tRNA, GTP and the small ribosomal subunit.

[1] Kyrpides NC, Woese CR, Proc Natl Acad Sci U S A 1998;95:3726-3730.

15 300. Initiation factor 3 signature

Initiation factor 3 (IF-3) (gene infC) [1] is one of the three factors required for the initiation of protein biosynthesis in bacteria. IF-3 is thought to function as a fidelity factor during the assembly of the ternary initiation complex which consist of the 30S ribosomal subunit, the initiator tRNA and the messenger RNA. IF-3 binds to the 30S ribosomal subunit; it is a basic protein of 141 to 212 residues. The chloroplast initiation factor IF-3(chl) is a protein that enhances the poly(A,U,G)-dependent binding of the initiator tRNA to chloroplast ribosomal30s subunits. In its mature form it is a protein of about 400 residues whose central section is evolutionary related to the sequence of bacterial IF-3 [2].As a signature pattern a highly conserved region was selected located in the central section of bacterial IF-3 and of 20 IF-3(chl).

25

Consensus pattern: [KR]-[LIVM](2)-[DN]-[FY]-[GSN]-[KR]-[LIVMFYS]-x-[FY]-[DEQTH]-x(2)-[KRQ]-

30 [1] Liveris D., Schwartz J.J., Geertman R., Schwartz I. FEMS Microbiol. Lett. 112:211-216(1993).

[2] Lin Q., Ma L., Burkhardt W., Spremulli L.L. J. Biol. Chem. 269:9436-9444(1994).

301. Imidazoleglycerol-phosphate dehydratase signatures (IGPD)

Imidazoleglycerol-phosphate dehydratase (EC 4.2.1.19) is the enzyme that catalyzes the seventh step in the biosynthesis of histidine in bacteria, fungi and plants. In most organisms it is a monofunctional protein of about 22 to 29 Kd. In some bacteria such as *Escherichia coli* it is the C-terminal domain of a bifunctional protein that includes a histidinol-phosphatase domain [1]. Two signature patterns were developed that each include two consecutive histidine residues.

Consensus pattern: [LIVMY]-[DE]-x-H-H-x(2)-E-x(2)-[GCA]-[LIVM]-[STAC]-[LIVM]-
Consensus pattern: G-x-[DN]-x-H-H-x(2)-E-[STAGC]-x-[FY]-K -

[1] Carlomagno M.S., Chiariotti L., Alifano P., Nappo A.G., Bruni C.B. J. Mol. Biol. 203:585-606(1988).

302. Indole-3-glycerol phosphate synthase signature (IGPS)

Indole-3-glycerol phosphate synthase (EC 4.1.1.48) (IGPS) catalyzes the fourth step in the biosynthesis of tryptophan: the ring closure of 1-(2-carboxy-phenylamino)-1-deoxyribulose into indol-3-glycerol-phosphate. In some bacteria, IGPS is a single chain enzyme. In others - such as *Escherichia coli* - it is the N-terminal domain of a bifunctional enzyme that also catalyzes N-(5'-phosphoribosyl)anthranilate isomerase (PRAI) activity, the third step of tryptophan biosynthesis. In fungi, IGPS is the central domain of a trifunctional enzyme that also contains a PRAI C-terminal domain and a glutamine amidotransferase N-terminal domain. The N-terminal section of IGPS contains a highly conserved region which X-ray crystallography studies [1] have shown to be part of the active site cavity. This region was used as a signature pattern for IGPS.

Consensus pattern: [LIVMFY]-[LIVMC]-x-E-[LIVMFYC]-K-[KRSP]-[STAK]-S-P-[ST]-
x(3)-[LIVMFYST]-

[1] Wilmanns M., Priestle J.P., Niermann T., Jansonius J.N. J. Mol. Biol. 223:477-507(1992).

303. (IL2) Interleukin 2. 31 members

304. (ILVD EDD) Dihydroxy-acid and 6-phosphogluconate dehydratases. Two dehydratases have been shown [1] to be evolutionary related: - Dihydroxy-acid dehydratase (EC 4.2.1.9) (gene *ilvD* or *ILV3*) which catalyzes the fourth step in the biosynthesis of isoleucine and valine, the dehydration of 2,3-dihydroxy-isovaleric acid into alpha-ketoisovaleric acid. - 6-phosphogluconate dehydratase (EC 4.2.1.12) (gene *edd*) which catalyzes the first step in the Entner-Doudoroff pathway, the dehydration of 6-phospho- D-gluconate into 6-phospho-2-dehydro-3-deoxy-D-gluconate. - Escherichia coli hypothetical protein *yjhG*. Both enzymes are proteins of about 600 amino acid residues. Two highly conserved regions have been developed as signature patterns. The first pattern is located in the N-terminal part and contains a cysteine that could be involved in the binding of a 2Fe-2S iron-sulfur cluster [2]. The second pattern is located in the C-terminal half.

Consensus pattern: C-D-K-x(2)-P-[GA]-x(3)-[GA] [The C could be a 2Fe-2S ligand]

Consensus pattern: [SA]-L-[LIVM]-T-D-[GA]-R-[LIVMF]-S-[GA]-[GAV]-[ST]-

[1] Egan S.E., Fliege R., Tong S., Shibata A., Wolf R.E. Jr., Conway T. J. Bacteriol. 174:4638-4646(1992).[2] Velasco J.A., Cansado J., Pena M.C., Kawakami T., Laborda J., Notario V. Gene 137:179-185(1993).

305. IMP dehydrogenase / GMP reductase signature

IMP dehydrogenase (EC 1.1.1.205) (IMPDH) catalyzes the rate-limiting reaction of de novo GTP biosynthesis, the NAD-dependent reduction of IMP into XMP [1]. Inhibition of IMP dehydrogenase activity results in the cessation of DNA synthesis. As IMP dehydrogenase is associated with cell proliferation, it is a possible target for cancer chemotherapy. Mammalian and bacterial IMPDHs are tetramers of identical chains. There are two IMP dehydrogenase isozymes in humans [2]. GMP reductase (EC 1.6.6.8) catalyzes the irreversible and NADPH-dependent reductive deamination of GMP into IMP [3]. It converts nucleobase, nucleoside

and nucleotide derivatives of G to A nucleotides, and maintains intracellular balance of A and G nucleotides. IMP dehydrogenase and GMP reductase share many regions of sequence similarity. One of these regions is centered on a cysteine residue thought [3] to be involved in binding IMP. This region was used as a signature pattern.

5

Consensus pattern: [LIVM]-[RK]-[LIVM]-G-[LIVM]-G-x-G-S-[LIVM]-C-x-T [C is the putative IMP-binding residue]-

[1] Collart F.R., Huberman E. J. Biol. Chem. 263:15769-15772(1988).

10 [2] Natsumeda Y., Ohno S., Kawasaki H., Konno Y., Weber G., Suzuki K. J. Biol. Chem. 265:5292-5295(1990).

[3] Andrews S.C., Guest J.R. Biochem. J. 255:35-43(1988).

15 306. (IPPC) Inositol polyphosphate phosphatase family, catalytic domain

[1] York JD, Ponder JW, Chen ZW, Mathews FS, Majerus PW; Biochemistry 1994;33:13164-13171. [2] Jefferson AB, Auethavekiat V, Pot DA, Williams LT, Majerus PW; J Biol Chem 1997;272:5983-5988. [3] Zhang X, Jefferson AB, 20 Auethavekiat V, Majerus PW; Proc Natl Acad Sci U S A 1995;92:4853-4856. [4] York JD, Majerus PW. Proc Natl Acad Sci U S A 1990;87:9548-9552. [5] Neuwald AF, York JD, Majerus PW; FEBS Lett 1991;294:16-18.

25

307. IQ calmodulin-binding motif

[1] Xie X, Harrison DH, Schlichting I, Sweet RM, Kalabokis VN, Szent-Gyorgyi AG, Cohen C; Nature 1994;368:306-312.

30 [2] Rhoads AR, Friedberg F; FASEB J 1997;11:331-340.

308. Inosine-uridine preferring nucleoside hydrolasefamily signature (IU nuc hydro)

Inosine-uridine preferring nucleoside hydrolase (EC 3.2.2.1) (IU-nucleosidehydrolase or IUNH) is an enzyme first identified in protozoan [1] that catalyzes the hydrolysis of all of the commonly occurring purine and pyrimidine nucleosides into ribose and the associated base, but has a preference for inosine and uridine as substrates. This enzyme is important for these parasitic organisms, which are deficient in de novo synthesis of purines, to salvage the host purine nucleosides. IUNH from *Crithidia fasciculata* has been sequenced and characterized, it is an homotetrameric enzyme of subunits of 34 Kd. An histidine has been shown to be important for the catalytic mechanism, it acts a proton donor to activate the hypoxanthine leaving group. IUNH is evolutionary related to a number of uncharacterized proteins from various biological sources, notably: - *Escherichia coli* hypothetical protein yaaF. - *Escherichia coli* hypothetical protein ybeK. - *Escherichia coli* hypothetical protein yeiK. - Fission yeast hypothetical protein SpAC17G8.02. - Yeast hypothetical protein YDR400w. - An hypothetical protein from the archaeobacteria *Desulfurolobus ambivalens*. As a signature pattern for these proteins, a highly conserved region was selected located in the N-terminal extremity. This region contains four conserved aspartates that have been shown [2] to be located in the active site cavity.

Consensus pattern: D-x-D-[PT]-[GA]-x-D-D-[TAV]-[VI]-A -

[1] Gopaul D.N., Meyer S.L., Degano M., Sacchettini J.C., Schramm V.L. *Biochemistry* 35:5963-5970(1996).

[2] Degano M., Gopaul D.N., Scapin G., Schramm V.L., Sacchettini J.C. *Biochemistry* 35:5971-5981(1996).

309. (Insulinase)

Insulinase family, zinc-binding region signature
(aka Peptidase_M16)

A number of proteases dependent on divalent cations for their activity have been shown [1,2] to belong to one family, on the basis of sequence similarity. These enzymes are listed below.

- Insulinase (EC 3.4.24.56) (also known as insulysin or insulin-degrading enzyme or IDE), a cytoplasmic enzyme which seems to be involved in the cellular processing of insulin, glucagon and other small polypeptides.

- Escherichia coli protease III (EC 3.4.24.55) (pitrilysin) (gene ptr), a periplasmic enzyme that degrades small peptides.

- Mitochondrial processing peptidase (EC 3.4.24.64) (MPP). This enzyme removes the transit peptide from the precursor form of proteins imported from the cytoplasm across the mitochondrial inner membrane. It is composed of two nonidentical homologous subunits termed alpha and beta. The beta subunit seems to be catalytically active while the alpha subunit has probably lost its activity.

- Nardilysin (EC 3.4.24.61) (N-arginine dibasic convertase or NRD convertase) this mammalian enzyme cleaves peptide substrates on the N-terminus of Arg residues in dibasic stretches.

- Klebsiella pneumoniae protein pqqF. This protein is required for the biosynthesis of the coenzyme pyrrolo-quinoline-quinone (PQQ). It is thought to be protease that cleaves peptide bonds in a small peptide (gene pqqA) thus providing the glutamate and tyrosine residues necessary for the synthesis of PQQ.

- Yeast protein AXL1, which is involved in axial budding [3].

- Eimeria bovis sporozoite developmental protein.

- Escherichia coli hypothetical protein yddC and HI1368, the corresponding Haemophilus influenzae protein.

- Bacillus subtilis hypothetical protein ymxG.

- Caenorhabditis elegans hypothetical proteins C28F5.4 and F56D2.1.

It should be noted that in addition to the above enzymes, this family also includes the core proteins I and II of the mitochondrial bc1 complex (also called cytochrome c reductase or complex III), but the situation as to the activity or lack of activity of these subunits is quite complex:

- In mammals and yeast, core proteins I and II lack enzymatic activity.

- In Neurospora crassa and in potato core protein I is equivalent to the beta subunit of MPP.

- In Euglena gracilis, core protein I seems to be active, while subunit II is inactive.

These proteins do not share many regions of sequence similarity; the most noticeable is in the N-terminal section. This region includes a conserved histidine followed, two residues later by a glutamate and another histidine. In pitrilysin, it has been shown [4] that this H-x-x-E-H motif is involved in enzyme activity; the two histidines bind zinc and the glutamate is necessary for catalytic activity. Non active members of this family have lost from one to three of these active site residues. We developed a signature pattern that detect active members of this family as well as some inactive members.

Consensus pattern G-x(8,9)-G-x-[STA]-H-[LIVMFY]-[LIVMC]-[DERN]-[HRKL]-[LMFAT]-x-[LFSTH]-x-[GSTAN]-[GST] [The two H are zinc ligands] [E is the active site residue] Sequences known to belong to this class detected by the pattern ALL active members as well as all MPP alpha subunits and core II subunits. Does not detect inactive core I subunits.

Note: these proteins belong to family M16 in the classification of peptidases [5].

□

[1] Rawlings N.D., Barrett A.J. Biochem. J. 275:389-391(1991).

□

[2] Braun H.-P., Schmitz U.K. Trends Biochem. Sci. 20:171-175(1995).

[3] Becker A.B., Roth R.A. Proc. Natl. Acad. Sci. U.S.A. 89:3835-3839(1992).

[4] Fujita A., Oka C., Arikawa Y., Katagai T., Tonouchi A., Kuhara S., Misumi Y. Nature 372:567-570(1994).

[5] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).

310. Involucrin repeat

Eckert RL, Yaffe MB, Crish JF, Murthy S, Rorke EA, Welter JF, J Invest Dermatol 1993;100:613-617.

311. Isochorismatase family. This family are hydrolase enzymes.

Romao MJ, Turk D, Gomis-Ruth FX, Huber R, Schumacher G, Mollering H, Russmann L, J Mol Biol 1992;226:1111-1130.

312. Inositol monophosphatase family signatures (inositol_P)

It has been shown [1] that several proteins share two sequence motifs. Two of these proteins are enzymes of the inositol phosphate second messenger signaling pathway: - Vertebrate and plants inositol monophosphatase (EC 3.1.3.25). - Vertebrate inositol polyphosphate 1-phosphatase (EC 3.1.3.57). The function of the other proteins is not yet clear: - Bacterial protein cysQ. CysQ could help to control the pool of PAPS (3'-phosphoadenoside 5'-phosphosulfate), or be useful in sulfite synthesis. - Escherichia coli protein suhB. Mutations in suhB results in the enhanced synthesis of heat shock sigma factor (htpR). - Neurospora crassa protein Qa-X. Probably involved in quinate metabolism. - Emericella nidulans protein qutG. Probably involved in quinate metabolism. - Yeast protein HAL2/MET22 [2] involved in salt tolerance as well as methionine biosynthesis. - Yeast hypothetical protein YHR046c. - Caenorhabditis elegans hypothetical protein F13G3.5. - A Rhizobium leguminosarum hypothetical protein encoded upstream of the pss gene for exopolysaccharide synthesis. - Methanococcus jannaschii hypothetical protein MJ0109. It is suggested [1] that these proteins may act by enhancing the synthesis or degradation of phosphorylated messenger molecules. From the X-ray structure of human inositol monophosphatase [3], it seems that some of the conserved residues are involved in binding a metal ion and/or the phosphate group of the substrate.

Consensus pattern: [FWV]-x(0,1)-[LIVM]-D-P-[LIVM]-D-[SG]-[ST]-x(2)-[FY]-x-

[HKRNSTY] [The first D and the T bind a metal ion]-

Consensus pattern: [WV]-D-x-[AC]-[GSA]-[GSAPV]-x-[LIVACP]-[LIV]-[LIVAC]-x(3)-[GH]-[GA]-

[1] Neuwald A.F., York J.D., Majerus P.W. FEBS Lett. 294:16-18(1991).

[2] Glaeser H.-U., Thomas D., Gaxiola R., Montrichard F., Surdin-Kerjan Y., Serrano R. EMBO J. 12:3105-3110(1993).

[3] Bone R., Springer J.P., Atack J.R. Proc. Natl. Acad. Sci. U.S.A. 89:10031-10035(1992).

313. Ion transport protein

This family contains Sodium, Potassium, Calcium ion channel. This family is 6 transmembrane helices in which the last two helices flank a loop which determines ion selectivity. In some sub-families (e.g. Na channels) the domain is repeated four times, whereas in others (e.g. K channels) the protein forms as a tetramer in the membrane. A bacterial structure of the protein is known for the last two helices but is not the Pfam family due to it lacking the first four helices.

314. Isocitrate and isopropylmalate dehydrogenases signature (isodh)

Isocitrate dehydrogenase (IDH) [1,2] is an important enzyme of carbohydrate metabolism which catalyzes the oxidative decarboxylation of isocitrate into alpha-ketoglutarate. IDH is either dependent on NAD⁺ (EC 1.1.1.41) or on NADP⁺ (EC 1.1.1.42). In eukaryotes there are at least three isozymes of IDH: two are located in the mitochondrial matrix (one NAD⁺-dependent, the other NADP⁺-dependent), while the third one (also NADP⁺-dependent) is cytoplasmic. In *Escherichia coli* the activity of a NADP⁺-dependent form of the enzyme is controlled by the phosphorylation of a serine residue; the phosphorylated form of IDH is completely inactivated. 3-isopropylmalate dehydrogenase (EC 1.1.1.85) (IMDH) [3,4] catalyzes the third step in the biosynthesis of leucine in bacteria and fungi, the oxidative decarboxylation of 3-isopropylmalate into 2-oxo-4-methylvalerate. Tartrate dehydrogenase (EC 1.1.1.93) [5] catalyzes the reduction of tartrate to oxaloglycolate. These enzymes are evolutionary related [1,3,4,5]. The best conserved region of these enzymes is a glycine-rich stretch of residues located in the C-terminal section. This region was used as a signature pattern.

Consensus pattern: [NS]-[LIMYT]-[FYDN]-G-[DNT]-[IMVY]-x-[STGDN]-[DN]-x(2)-[SGAP]-x(3,4)-G-[STG]-[LIVMPA]-G-[LIVMF]-

[1] Hurley J.H., Thorsness P.E., Ramalingam V., Helmers N.H., Koshland D.E. Jr., Stroud R.M. Proc. Natl. Acad. Sci. U.S.A. 86:8635-8639(1989).

[2] Cupp J.R., McAlister-Henn L. J. Biol. Chem. 266:22199-22205(1991).

[3] Imada K., Sato M., Tanaka N., Katsube Y., Matsuura Y., Oshima T. J. Mol. Biol. 222:725-738(1991).

[4] Zhang T., Koshland D.E. Jr. Protein Sci. 4:84-92(1995).

[5] Tipton P.A., Beecher B.S. Arch. Biochem. Biophys. 313:15-21(1994).

5

315. Jacalin-like lectin domain.

Proteins containing this domain are lectins. It is found in

10 1 to 6 copies in these proteins. The domain is also found in the animal prostatic spermine-binding protein ([Swiss:P15501](#)).

[1] Sankaranarayanan R, Sekar K, Banerjee R, Sharma V, Surolia A, Vijayan M; Nat Struct Biol 1996;3:596-603.

316. KH domain

KH motifs probably bind RNA directly. Auto antibodies to Nova, a KH domain protein, cause paraneoplastic opsoclonus ataxia.

[1] Burd CG, Dreyfuss G, Science 1994;265:615-621.

[2] Musco G, Stier G, Joseph C, Castiglione Morelli MA, Nilges M, Gibson TJ, Pastore A, Cell 1996;85:237-245.

25 317. Kelch motif

The kelch motif was initially discovered in Kelch ([Swiss:Q04652](#)). In this protein there are six copies of the motif. It has been shown that [Swiss:Q04652](#) is related to Galactose Oxidase [1] for which a structure has been solved [2]. The kelch motif forms a beta sheet. Several of these sheets associate to form a beta propeller structure as found in [neur](#),

30 [1] Bork P, Doolittle RF, J Mol Biol 1994;236:1277-1282. [2] Ito N, Phillips SE, Stevens C, Ogel ZB, McPherson MJ, Keen, JN, Yadav KD, Knowles PF, Nature 1991;350:87-90.

318. Soybean trypsin inhibitor (Kunitz) protease inhibitors family signature

The soybean trypsin inhibitor (Kunitz) family [1] is one of the numerous families of proteinase inhibitors. It comprise plant proteins which have inhibitory activity against serine proteinases from the trypsin and subtilisin families, thiol proteinases and aspartic proteinases as well as some proteins that are probably involved in seed storage. This family is currently known to group the following proteins: - Trypsin inhibitors A, B, C, KTI1, and KTI2 from soybean. - Trypsin inhibitor DE3 from coral beans (*Erythrina* sp.). - Trypsin inhibitor DE5 from sandal bead tree. - Trypsin inhibitors 1A (WTI-1A), 1B (WTI-1B), and 2 (WTI-2) from goa bean. - Trypsin inhibitor from *Acacia confusa*. - Trypsin inhibitor from silk tree. - Chymotrypsin inhibitor 3 (WCI-3) from goa bean. - Cathepsin D inhibitors PDI and NDI from potato [2], which inhibit both cathepsin D (aspartic proteinase) and trypsin. - Alpha-amylase/subtilisin inhibitors from barley and wheat. - Albumin-1 (WBA-1) from goa bean seeds [3]. - Miraculin from *Richadella dulcifica* [4], a sweet taste protein. - Sporamin from sweet potato [5], the major tuberous root protein. - Thiol proteinase inhibitor PCPI 8.3 (P340) from potato tuber [6]. - Wound responsive protein gwin3 from poplar tree [7]. - 21 Kd seed protein from cocoa [8]. All these proteins contain from 170 to 200 amino acid residues and one or two intrachain disulfide bonds. The best conserved region is found in their N-terminal section and is used as a signature pattern

Consensus pattern: [LIVM]-x-D-x-[EDNTY]-[DG]-[RKHDENQ]-x-[LIVM]-x(5)-Y-x-[LIVM] -

[1] Laskowski M., Kato I. *Annu. Rev. Biochem.* 49:593-626(1980).

[2] Ritonja A., Krizaj I., Mesko P., Kopitar M., Lucovnik P., Strukelj B., Pungercar J., Buttler D.J., Barrett A.J., Turk V. *FEBS Lett.* 267:13-15(1990).

[3] Kortt A.A., Strike P.M., de Jersey J. *Eur. J. Biochem.* 181:403-408(1989).

[4] Theerasilp S., Hitotsuya H., Nakajo S., Nakaja K., Nakamura Y., Kurihara Y. *J. Biol. Chem.* 264:6655-6659(1989).

[5] Hattori T., Yoshida N., Nakamura K. *Plant Mol. Biol.* 13:563-572(1989).

[6] Krizaj I., Drobic-Kosorok M., Brzin J., Jerala R., Turk V. *FEBS Lett.* 333:15-20(1993).

[7] Bradshaw H.D., Hollick J.B., Parsons T.J., Clarke H.R.G., Gordon M.P. *Plant Mol. Biol.* 14:51-59(1989).

[8] Tai H., McHenry L., Fritz P.J., Furtek D.B. Plant Mol. Biol. 16:913-915(1991).

319. Beta-ketoacyl synthases active site

Beta-ketoacyl-ACP synthase (KAS) [1] is the enzyme that catalyzes the condensation of malonyl-ACP with the growing fatty acid chain. It is found as a component of the following enzymatic systems: - Fatty acid synthetase (FAS), which catalyzes the formation of long-chain fatty acids from acetyl-CoA, malonyl-CoA and NADPH. Bacterial and plant chloroplast FAS are composed of eight separate subunits which correspond to different enzymatic activities; beta-ketoacyl synthase is one of these polypeptides. Fungal FAS consists of two multifunctional proteins, FAS1 and FAS2; the beta-ketoacyl synthase domain is located in the C-terminal section of FAS2. Vertebrate FAS consists of a single multifunctional chain; the beta-ketoacyl synthase domain is located in the N-terminal section [2]. - The multifunctional 6-methylsalicylic acid synthase (MSAS) from *Penicillium patulum* [3]. This is a multifunctional enzyme involved in the biosynthesis of a polyketide antibiotic and which has a KAS domain in its N-terminal section. - Polyketide antibiotic synthase enzyme systems. Polyketides are secondary metabolites produced by microorganisms and plants from simple fatty acids. KAS is one of the components involved in the biosynthesis of the *Streptomyces* polyketide antibiotics granatacin [4], tetracenomyacin C [5] and erythromycin. - *Emericella nidulans* multifunctional protein Wa. Wa is involved in the biosynthesis of conidial green pigment. Wa is protein of 216 Kd that contains a KAS domain. - *Rhizobium* nodulation protein nodeE, which probably acts as a beta-ketoacyl synthase in the synthesis of the nodulation Nod factor fatty acyl chain. - Yeast mitochondrial protein CEM1. The condensation reaction is a two step process: the acyl component of an activated acyl primer is transferred to a cysteine residue of the enzyme and is then condensed with an activated malonyl donor with the concomitant release of carbon dioxide. The sequence around the active site cysteine is well conserved and can be used as a signature pattern.

Consensus pattern: G-x(4)-[LIVMFAP]-x(2)-[AGC]-C-[STA](2)-[STAG]-x(3)-[LIVMF] [C is the active site residue]

[1] Kauppinen S., Siggaard-Andersen M., von Wettstein-Knowles P. Carlsberg Res. Commun. 53:357-370(1988).

[2] Witkowski A., Rangan V.S., Randhawa Z.I., Amy C.M., Smith S. Eur. J. Biochem. 198:571-579(1991).

[3] Beck J., Ripka S., Siegner A., Schiltz E., Schweizer E. Eur. J. Biochem. 192:487-498(1990).

5 [4] Bibb M.J., Biro S., Motamedi H., Collins J.F., Hutchinson C.R. EMBO J. 8:2727-2736(1989).

[5] Sherman D.H., Malpartida F., Bibb M.J., Kieser H.M., Bibb M.J., Hopwood D.A. EMBO J. 8:2717-2725(1989).

10

320. Kinesin motor domain signature and profile

Kinesin [1,2,3] is a microtubule-associated force-producing protein that may play a role in organelle transport. Kinesin is an oligomeric complex composed of two heavy chains and two light chains. The kinesin motor activity is directed toward the microtubule's plus end. The heavy chain is composed of three structural domains: a large globular N-terminal domain which is responsible for the motor activity of kinesin (it is known to hydrolyze ATP, to bind and move on microtubules), a central alpha-helical coiled coil domain that mediates the heavy chain dimerization; and a small globular C-terminal domain which interacts with other proteins (such as the kinesin light chains), vesicles and membranous organelles. A number of proteins have been recently found that contain a domain similar to that of the kinesin 'motor' domain [1,4,E1]: - *Drosophila* claret segregational protein (ncd). Ncd is required for normal chromosomal segregation in meiosis, in females, and in early mitotic divisions of the embryo. The ncd motor activity is directed toward the microtubule's minus end. - *Drosophila* kinesin-like protein (nod). Nod is required for the distributive chromosome segregation of nonexchange chromosomes during meiosis. - Human CENP-E [4]. CENP-E is a protein that associates with kinetochores during chromosome congression, relocates to the spindle midzone at anaphase, and is quantitatively discarded at the end of the cell division. CENP-E is probably an important motor molecule in chromosome movement and/ or spindle elongation. - Human mitotic kinesin-like protein-1 (MKLP-1), a motor protein whose activity is directed toward the microtubule's plus end. - Yeast KAR3 protein, which is essential for yeast nuclear fusion during mating. KAR3 may mediate microtubule sliding during nuclear fusion and possibly mitosis. - Yeast CIN8 and KIP1 proteins which are required for the assembly of the mitotic spindle. Both proteins seem to interact with spindle microtubules to

30

produce an outwardly directed force acting upon the poles. - Fission yeast cut7 protein, which is essential for spindle body duplication during mitotic division. - *Emericella nidulans* bimC, which plays an important role in nuclear division. - *Emericella nidulans* klpA. - *Caenorhabditis elegans* unc-104, which may be required for the transport of substances needed for neuronal cell differentiation. - *Caenorhabditis elegans* osm-3. - *Xenopus* Eg5, which may be involved in mitosis. - *Arabidopsis thaliana* KatA, KatB and katC. - *Chlamydomonas reinhardtii* FLA10/KHP1 and KLP1. Both proteins seem to play a role in the rotation or twisting of the microtubules of the flagella. - *Caenorhabditis elegans* hypothetical protein T09A5.2. The kinesin motor domain is located in the N-terminal part of most of the above proteins, with the exception of KAR3, klpA, and ncd where it is located in the C-terminal section. The kinesin motor domain contains about 330 amino acids. An ATP-binding motif of type A is found near position 80 to 90, the C-terminal half of the domain is involved in microtubule-binding. The signature pattern for that domain is derived from a conserved decapeptide inside the microtubule-binding part.

Consensus pattern: [GSA]-[KRHPSTQVM]-[LIVMF]-x-[LIVMF]-[IVC]-D-L-[AH]-G-[SAN]-E

[1] Bloom G.S., Endow S.A. Protein Prof. 2:1109-1171(1995).

[2] Vallee R.B., Shpetner H.S. Annu. Rev. Biochem. 59:909-932(1990).

[3] Brady S.T. Trends Cell Biol. 5:159-164(1995).

[4] Endow S.A. Trends Biochem. Sci. 16:221-225(1991).[E1]

321. Ribosomal protein L15 signature

Ribosomal protein L15 is one of the proteins from the large ribosomal subunit. In *Escherichia coli*, L15 is known to bind the 23S rRNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups: - Eubacterial L15. - Plant chloroplast L15 (nuclear-encoded). - Archaeobacterial L15. - Vertebrate L27a. - *Tetrahymena thermophila* L29. - Fungi L27a (L29, CRP-1, CYH2). L15 is a protein of 144 to 154 amino-acid residues. As a signature pattern, a conserved region was selected in the C-terminal section of these proteins.

323

Consensus pattern: K-[LIVM](2)-[GASL]-x-[GT]-x-[LIVMA]-x(2,5)-[LIVM]-x- [LIVMF]-x(3,4)-[LIVMFCA]-[ST]-x(2)-A-x(3)-[LIVM]-x(3)-G

[1] Otake E., Hashimoto T., Mizuta K., Suzuki K. Protein Seq. Data Anal. 5:301-313(1993).

5

322. LBP / BPI / CETP family signature

The following mammalian lipid-binding serum glycoproteins belong to the same family [1,2,3]: - Lipopolysaccharide-binding protein (LBP). LBP binds to the lipid A moiety of bacterial lipopolysaccharides (LPS), a glycolipid present in the outer membrane of all Gram-negative bacteria. The LBP/LPS complex seems to interact with the CD14 receptor and may be responsible for the secretion of alpha-TNF. - Bactericidal permeability-increasing protein (BPI). Like LBP, BPI binds LPS and has a cytotoxic activity on Gram-negative bacteria. - Cholesteryl ester transfer protein (CETP). CETP is involved in the transfer of insoluble cholesteryl esters in reverse cholesterol transport. - Phospholipid transfer protein (PLTP). May play a key role in extracellular phospholipid transport and modulation of HDL particles. These proteins are structurally related and share many regions of sequence similarities. As a signature pattern one of these regions was selected, which is located in the N-terminal section of these proteins; a region which could be involved in the binding to the lipids [2].

Consensus pattern: [PA]-[GA]-[LIVMC]-x(2)-R-[IV]-[ST]-x(3)-L-x(5)-[EQ]-x(4)- [LIVM]-[EQK]-x(8)-P

[1] Schumann R.R., Leong S.R., Flaggs G.W., Gray P.W., Wright S.D., Mathison J.C., Tobias P.S., Ulevitch R.J. Science 249:1429-1431(1990).

[2] Gray P.W., Flaggs G., Leong S.R., Gumina R.J., Weiss J., Ooi C.E., Elsbach P. J. Biol. Chem. 264:9505-9509(1989).

[3] Day J.R., Albers J.J., Lofton-Day C.E., Gilbert T.L., Ching A.F.T., Grant F.J., O'Hara P.J., Marcovina S.M., Adolphson J.L. J. Biol. Chem. 269:9388-9391(1994).

323. LIM domain signature and profile

Recently [1,2] a number of proteins have been found to contain a conserved cysteine-rich domain of about 60 amino-acid residues. These proteins are: - *Caenorhabditis elegans* mec-3; a protein required for the differentiation of the set of six touch receptor neurons in this nematode. - *Caenorhabditis elegans* lin-11; a protein required for the asymmetric division of vulval blast cells. - Vertebrate insulin gene enhancer binding protein isl-1. Isl-1 binds to one of the two cis-acting protein-binding domains of the insulin gene. - Vertebrate homeobox proteins lim-1, lim-2 (lim-5) and lim3. - Vertebrate lmx-1, which acts as a transcriptional activator by binding to the FLAT element; a beta-cell-specific transcriptional enhancer found in the insulin gene. - Mammalian LH-2, a transcriptional regulatory protein involved in the control of cell differentiation in developing lymphoid and neural cell types. - *Drosophila* protein apterous, required for the normal development of the wing and halter imaginal discs. - Vertebrate protein kinases LIMK-1 and LIMK-2. - Mammalian rhombotins. Rhombotin 1 (RBTN1 or TTG-1) and rhombotin-2 (RBTN2 or TTG-2) are proteins of about 160 amino acids whose genes are disrupted by chromosomal translocations in T-cell leukemia. - Mammalian and avian cysteine-rich protein (CRP), a 192 amino-acid protein of unknown function. Seems to interact with zyxin. - Mammalian cysteine-rich intestinal protein (CRIP), a small protein which seems to have a role in zinc absorption and may function as an intracellular zinc transport protein. - Vertebrate paxillin, a cytoskeletal focal adhesion protein. - Mouse testin. Mouse testin should not be confused with rat testin which is a thiol protease homolog. - Sunflower pollen specific protein SF3. - Chicken zyxin. Zyxin is a low-abundance adhesion plaque protein which has been shown to interact with CRP. - Yeast protein LRG1 which is involved in sporulation [4]. - Yeast rho-type GTPase activating protein RGA1/DBM1. - *Caenorhabditis elegans* homeobox protein ceh-14. - *Caenorhabditis elegans* homeobox protein unc-97. - Yeast hypothetical protein YKR090w. - *Caenorhabditis elegans* hypothetical proteins C28H8.6. These proteins generally have two tandem copies of a domain, called LIM (for Lin-11 Isl-1 Mec-3) in their N-terminal section. Zyxin and paxillin are exceptions in that they contain respectively three and four LIM domains at their C-terminal extremity. In apterous, isl-1, LH-2, lin-11, lim-1 to lim-3, lmx-1 and ceh-14 and mec-3 there is a homeobox domain some 50 to 95 amino acids after the LIM domains. In the LIM domain, there are seven conserved cysteine residues and a histidine. The arrangement followed by these conserved residues is C-x(2)-C-x(16,23)-H-x(2)-[CH]-x(2)-C-x(2)-C-x(16,21)-C-x(2,3)-[CHD]. The LIM domain binds two zinc ions [5]. LIM does not bind DNA,

rather it seems to act as interface for protein-protein interaction. A pattern was developed that spans the first half of the LIM domain.

Consensus pattern: C-x(2)-C-x(15,21)-[FYWH]-H-x(2)-[CH]-x(2)-C-x(2)-C-x(3)- [LIVMF]
[The 5 C's and the H bind zinc]

[1] Freyd G., Kim S.K., Horvitz H.R. Nature 344:876-879(1990).

[2] Baltz R., Evrard J.-L., Domon C., Steinmetz A. Plant Cell 4:1465-1466(1992).

[3] Sanchez-Garcia I., Rabbitts T.H. Trends Genet. 10:315-320(1994).

[4] Mueller A., Xu G., Wells R., Hollenberg C.P., Piepersberg W. Nucleic Acids Res. 22:3151-3154(1994).

[5] Michelsen J.W., Schmeichel K.L., Beckerle M.C., Winge D.R. Proc. Natl. Acad. Sci. U.S.A. 90:4404-4408(1993).

324. (LRR) Leucine Rich Repeat

CAUTION: This Pfam may not find all Leucine Rich Repeats in a protein. Leucine Rich Repeats are short sequence motifs present in a number of proteins with diverse functions and cellular locations. These repeats are usually involved in protein-protein interactions. Each Leucine Rich Repeat is composed of a beta-alpha unit. These units form elongated non-globular structures. Leucine Rich Repeats are often flanked by cysteine rich domains.

Number of members: 3017

[1] The leucine-rich repeat: a versatile binding motif. Kobe B, Deisenhofer J; Trends Biochem Sci 1994;19:415-421. [2] Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. Kobe B, Deisenhofer J; Nature 1993;366:751-756.

325. Plant lipid transfer protein family signature (LTP)

Plant cells contain proteins, called lipid transfer proteins (LTP) [1,2,3], which are able to facilitate the transfer of phospholipids and other lipids across membranes. These proteins, whose subcellular location is not yet known, could play a major role in membrane biogenesis by conveying phospholipids such as waxes or cutin from their site of biosynthesis to membranes unable to form these lipids. Plant LTP's are proteins of about 9 Kd (90 amino

acids) which contain eight conserved cysteine residues all involved in disulfide bridges, as shown in the following schematic representation.

```

+-----+ | +-----+ |||| *****
xCxxxxCxxxxxxCCxxxxxxxCxCxxxxxxxxxxCxxxxxCxx ||| +-----|-----+ | +---
5 -----+

```

'C': conserved cysteine involved in a disulfide bond.

'*': position of the pattern.

Consensus pattern: [LIVM]-[PA]-x(2)-C-x-[LIVM]-x-[LIVM]-x-[LIVMFY]-x-[LIVM]-
10 [ST]-x(3)-[DN]-C-x(2)-[LIVM] [The two C's are involved in disulfide bonds]

[1] Wirtz K.W.A. Annu. Rev. Biochem. 60:73-99(1991).

[2] Arondel V., Kader J.C. Experientia 46:579-585(1990).

[3] Ohlrogge J.B., Browse J., Somerville C.R. Biochim. Biophys. Acta 1082:1-26(1991).

326. (LAMP) Lysosome-associated membrane glycoproteins signatures

Lysosome-associated membrane glycoproteins (lamp) [1] are integral membrane proteins, specific to lysosomes, and whose exact biological function is not yet clear. Structurally, the
20 lamp proteins consist of two internally homologous lysosome-luminal domains separated by a proline-rich hinge region; at the C-terminal extremity there is a transmembrane region followed by a very short cytoplasmic tail. In each of the duplicated domains, there are two conserved disulfide bonds. This structure is schematically represented in the figure below. +--

```

---+ +-----+ +-----+ +-----+ |||||

```

```

25 xCxxxxxCxxxxxxxxxxxCxxxxxCxxxxxxxxxCxxxxxCxxxxxxxxxxxCxxxxxCxxxxxxxxx <-
-----><Hinge><-----><TM><C>

```

In mammals, there are two closely related types of lamp: lamp-1 and lamp-2. In chicken lamp-1 is known as LEP100. The macrophage protein CD68 (or macrosialin) [2] is a heavily glycosylated integral membrane protein whose structure consists of a mucin-like domain followed by a proline-rich
30 hinge; a single lamp-like domain; a transmembrane region and a short cytoplasmic tail. Two signature patterns for this family of proteins were developed. The first one is centered on the first conserved cysteine of the duplicated domains. The second corresponds to a region that

includes the extremity of the second domain, the totality of the transmembrane region and the cytoplasmic tail.

Consensus pattern: [STA]-C-[LIVM]-[LIVMFYW]-A-x-[LIVMFYW]-x(3)-[LIVMFYW]-
x(3)-Y [C is involved in a disulfide bond] –

Consensus pattern: C-x(2)-D-x(3,4)-[LIVM](2)-P-[LIVM]-x-[LIVM]-G-x(2)-[LIVM]- x-G-
[LIVM](2)-x-[LIVM](4)-A-[FY]-x-[LIVM]-x(2)-[KR]-[RH]- x(1,2)-[STAG](2)-Y-[EQ] [C
is involved in a disulfide bond]

[1] Fukuda M. J. Biol. Chem. 266:21327-21330(1991).

[2] Holness C.L., da Silva R.P., Fawcett J., Gordon S., Simmons D.L. J. Biol. Chem.
268:9661-9666(1993).

327. Lipolytic enzymes "G-D-S-L" family, serine active site

Recently [1], a family of lipolytic enzymes has been characterized. This family currently consist of the following proteins:

- Aeromonas hydrophila lipase/phosphatidylcholine-sterol acyltransferase.
- Xenorhabdus luminescens lipase 1.
- Vibrio mimicus arylesterase.
- Escherichia coli acyl-coA thioesterase I (gene tesA).
- Vibrio parahaemolyticus thermolabile hemolysin/atypical phospholipase.
- Rabbit phospholipase AdRab-B, an intestinal brush border protein with esterase and phospholipase A/lysophospholipase activity that could be involved in the uptake of dietary lipids. AdRab-B contains four repeats of about 320 amino acids.
- Arabidopsis thaliana and Brassica napus anther-specific proline-rich protein APG.
- A Pseudomonas putida hypothetical protein in trpE-trpG intergenic region. A serine has been identified a part of the active site in the Aeromonas, Vibrio mimicus and Escherichia coli enzymes. It is located in a conserved sequence motif that can be used as a signature pattern for these proteins.

-Consensus pattern: [LIVMFYAG](4)-G-D-S-[LIVM]-x(1,2)-[TAG]-G
[S is the active site residue]

328. (Lipoprotein 4) Prokaryotic membrane lipoprotein lipid attachment site

In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signalpeptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]): - Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp). - *Escherichia coli* lipoprotein-28 (gene nlpA). - *Escherichia coli* lipoprotein-34 (gene nlpB). - *Escherichia coli* lipoprotein nlpC. - *Escherichia coli* lipoprotein nlpD. - *Escherichia coli* osmotically inducible lipoprotein B (gene osmB). - *Escherichia coli* osmotically inducible lipoprotein E (gene osmE). - *Escherichia coli* peptidoglycan-associated lipoprotein (gene pal). - *Escherichia coli* rare lipoproteins A and B (genes rplA and rplB). - *Escherichia coli* copper homeostasis protein cutF (or nlpE). - *Escherichia coli* plasmids traT proteins. - *Escherichia coli* Col plasmids lysis proteins. - A number of *Bacillus* beta-lactamases. - *Bacillus subtilis* periplasmic oligopeptide-binding protein (gene oppA). - *Borrelia burgdorferi* outer surface proteins A and B (genes ospA and ospB). - *Borrelia hermsii* variable major protein 21 (gene vmp21) and 7 (gene vmp7). - *Chlamydia trachomatis* outer membrane protein 3 (gene omp3). - *Fibrobacter succinogenes* endoglucanase cel-3. - *Haemophilus influenzae* proteins Pal and Pcp. - *Klebsiella pullulunase* (gene pula). - *Klebsiella pullulunase* secretion protein pulS. - *Mycoplasma hyorhinis* protein p37. - *Mycoplasma hyorhinis* variant surface antigens A, B, and C (genes vlpABC). - *Neisseria* outer membrane protein H.8. - *Pseudomonas aeruginosa* lipopeptide (gene lppL). - *Pseudomonas solanacearum* endoglucanase egl. - *Rhodopseudomonas viridis* reaction center cytochrome subunit (gene cytC). - *Rickettsia* 17 Kd antigen. - *Shigella flexneri* invasion plasmid proteins mxiJ and mxiM. - *Streptococcus pneumoniae* oligopeptide transport protein A (gene amiA). - *Treponema pallidum* 34 Kd antigen. - *Treponema pallidum* membrane protein A (gene tmpA). - *Vibrio harveyi* chitinase (gene chb). - *Yersinia* virulence plasmid protein yscJ. - Halocyanin from *Natrobacterium pharaonis* [4], a membrane associated copper-binding protein. This is the first archaeobacterial protein known to be modified in such a fashion). From the precursor sequences of all these proteins, a consensus pattern and a set of rules to identify this type of post-translational modification was derived.

Consensus pattern: {DERK}(6)-[LIVMFIRSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1) The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least one Lys or one Arg in the first seven positions of the sequence.

5

[1] Hayashi S., Wu H.C. J. Bioenerg. Biomembr. 22:451-471(1990).

[2] Klein P., Somorjai R.L., Lau P.C.K. Protein Eng. 2:15-20(1988).

[3] von Heijne G. Protein Eng. 2:531-534(1989).

[4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol.

10 Chem. 269:14939-14945(1994).

329. (Lipoprotein 5) Prokaryotic membrane lipoprotein lipid attachment site. In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]): - Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp). - Escherichia coli lipoprotein-28 (gene nlpA). - Escherichia coli lipoprotein-34 (gene nlpB). - Escherichia coli lipoprotein nlpC. - Escherichia coli lipoprotein nlpD. - Escherichia coli osmotically inducible lipoprotein B (gene osmB). - Escherichia coli osmotically inducible lipoprotein E (gene osmE). - Escherichia coli peptidoglycan-associated lipoprotein (gene pal). - Escherichia coli rare lipoproteins A and B (genes rplA and rplB). - Escherichia coli copper homeostasis protein cutF (or nlpE). - Escherichia coli plasmids traT proteins. - Escherichia coli Col plasmids lysis proteins. - A number of Bacillus beta-lactamases. - Bacillus subtilis periplasmic oligopeptide-binding protein (gene oppA). - Borrelia burgdorferi outer surface proteins A and B (genes ospA and ospB). - Borrelia hermsii variable major protein 21 (gene vmp21) and 7 (gene vmp7). - Chlamydia trachomatis outer membrane protein 3 (gene omp3). - Fibrobacter succinogenes endoglucanase cel-3. - Haemophilus influenzae proteins Pal and Pcp. - Klebsiella pullulunase (gene pulA). - Klebsiella pullulunase secretion protein pulS. - Mycoplasma hyorhinis protein p37. - Mycoplasma hyorhinis variant surface antigens A, B, and C (genes vlp ABC). - Neisseria outer membrane protein H.8. - Pseudomonas aeruginosa lipopeptide (gene lppL). -

25

30

Pseudomonas solanacearum endoglucanase egl. - *Rhodopseudomonas viridis* reaction center cytochrome subunit (gene *cytC*). - *Rickettsia* 17 Kd antigen. - *Shigella flexneri* invasion plasmid proteins *mxjJ* and *mxjM*. - *Streptococcus pneumoniae* oligopeptide transport protein A (gene *amiA*). - *Treponema pallidum* 34 Kd antigen. - *Treponema pallidum* membrane protein A (gene *tmpA*). - *Vibrio harveyi* chitinase (gene *chb*). - *Yersinia* virulence plasmid protein *yscJ*. - Halocyanin from *Natrobacterium pharaonis* [4], a membrane associated copper-binding protein. This is the first archaeobacterial protein known to be modified in such a fashion). From the precursor sequences of all these proteins, a consensus pattern and a set of rules to identify this type of post-translational modification have been developed.

Consensus pattern: {DERK}(6)-[LIVMFIRSTAG](2)-[LIVMFIRSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1) The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least one Lys or one Arg in the first seven positions of the sequence.

[1] Hayashi S., Wu H.C. J. Bioenerg. Biomembr. 22:451-471(1990).[2] Klein P., Somorjai R.L., Lau P.C.K. Protein Eng. 2:15-20(1988).[3] von Heijne G. Protein Eng. 2:531-534(1989).[4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).

330. (Lum binding) Riboflavin synthase alpha chain family Lum-binding site signature
The following proteins have been shown [1,2] to be structurally and evolutionary related: -
Riboflavin synthase alpha chain (RS-alpha) (gene *ribC* in *Escherichia coli*, *ribB* in *Bacillus subtilis* and *Photobacterium leiognathi*, RIB5 in yeast). This enzyme synthesizes riboflavin from two moles of 6,7- dimethyl-8-(1'-D-ribityl)lumazine (Lum), a pteridine-derivative. -
Photobacterium phosphoreum lumazine protein (LumP) (gene *luxL*). LumP is a protein that modulates the color of the bioluminescence emission of bacterial luciferase. In the presence of LumP, light emission is shifted to higher energy values (shorter wavelength). LumP binds non-covalently to 6,7-dimethyl-8-(1'-D-ribityl) lumazine. - *Vibrio fischeri* yellow fluorescent protein (YFP) (gene *luxY*). Like LumP, YFP modulates light emission but towards a longer wavelength. YFP binds non-covalently to FMN. These proteins seem to have evolved from the duplication of a domain of about 100 residues. In its C-terminal section, this domain

331

contains a conserved motif [KR]-V-N-[LI]-E which has been proposed to be the binding site for Lum.RS-alpha which binds two molecules of Lum has two perfect copies of this motif, while LumP which binds one molecule of Lum, has a Glu instead of Lys/Arg in the first position of the second copy of the motif. Similarly, YFP, which binds to one molecule of FMN, also seems to have a potentially dysfunctional binding site by substitution of Gly for Glu in the last position of the first copy of the motif. Our signature pattern includes the Lum-binding motif.

Consensus pattern: [LIVMF]-x(5)-G-[STADNQ]-[KREQIYW]-V-N-[LIVM]-E

[1] O'Kane D.J., Woodward B., Lee J., Prasher D.C. Proc. Natl. Acad. Sci. U.S.A. 88:1100-1104(1991).

[2] O'Kane D.J., Prasher D.C. Mol. Microbiol. 6:443-449(1992).

331. Lysyl oxidase putative copper-binding region signature

Lysyl oxidase (LOX) [1] is an extracellular copper-dependent enzyme that catalyzes the oxidative deamination of peptidyl lysine residues in precursors of various collagens and elastins. The deaminated lysines are then able to form aldehyde cross-links. LOX binds a single copper atom which seems to reside within an octahedral coordination complex which includes at least three histidine ligands. Four histidine residues are clustered in a central region of the enzyme. This region is thought to be involved in copper-binding and is called the 'copper-talon' [1]. This region was used as a signature pattern.

Consensus pattern: W-E-W-H-S-C-H-Q-H-Y-H

[1] Krebs C.J., Krawetz S.A. Biochim. Biophys. Acta 1202:7-12(1993).

332. Metallo-beta-lactamase superfamily (lactamase_B)

[1] : Neuwald AF, Liu JS, Lipman DJ, Lawrence CE, Nucleic Acids Res 1997;25:1665-1677. [2] Carfi A, Pares S, Duee E, Galleni M, Duez C, Frere JM, Dideberg O, EMBO J 1995;14:4914-4921.

333. L-lactate dehydrogenase active site (ldh1)

L-lactate dehydrogenase (EC 1.1.1.27) (LDH) [1] catalyzes the reversible NAD-dependent interconversion of pyruvate to L-lactate. In vertebrate muscles and in lactic acid bacteria it represents the final step in anaerobic glycolysis. This tetrameric enzyme is present in prokaryotic and eukaryotic organisms. Invertebrates there are three isozymes of LDH: the M form (LDH-A), found predominantly in muscle tissues; the H form (LDH-B), found in heart muscle and the X form (LDH-C), found only in the spermatozoa of mammals and birds. In birds and crocodilian eye lenses, LDH-B serves as a structural protein and is known as epsilon-crystallin [2]. L-2-hydroxyisocaproate dehydrogenase (EC 1.1.1.-) (L-hicDH) [3] catalyzes the reversible and stereospecific interconversion between 2-ketocarboxylic acids and L-2-hydroxy-carboxylic acids. L-hicDH is evolutionary related to LDH's. As a signature for LDH's a region was selected that includes a conserved histidine which is essential to the catalytic mechanism.

Consensus pattern: [LIVMA]-G-[EQ]-H-G-[DN]-[ST] [H is the active site residue] -

[1] Abad-Zapatero C., Griffith J.P., Sussman J.L., Rossmann M.G. J. Mol. Biol. 198:445-467(1987).

[2] Hendriks W., Mulders J.W.M., Bibby M.A., Slingsby C., Bloemendal H., de Jong W.W. Proc. Natl. Acad. Sci. U.S.A. 85:7114-7118(1988).

[3] Lerch H.-P., Frank R., Collins J. Gene 83:263-270(1989).

25 Malate dehydrogenase active site signature (ldh2)

Malate dehydrogenase (EC 1.1.1.37) (MDH) [1,2] catalyzes the interconversion of malate to oxaloacetate utilizing the NAD/NADH cofactor system. The enzyme participates in the citric acid cycle and exists in all aerobic organisms. While prokaryotic organisms contains a single form of MDH, in eukaryotic cells there are two isozymes: one which is located in the mitochondrial matrix and the other in the cytoplasm. Fungi and plants also harbor a glyoxysomal form which functions in the glyoxylate pathway. In plants chloroplast there is an additional NADP-dependent form of MDH (EC 1.1.1.82) which is essential for both the universal C3 photosynthesis (Calvin) cycle and the more specialized C4 cycle. As a signature

pattern for this enzyme a region was chosen that includes two residues involved in the catalytic mechanism [3]: an aspartic acid which is involved in a proton relay mechanism, and an arginine which binds the substrate.

5 Consensus pattern: [LIVM]-T-[TRKMN]-L-D-x(2)-R-[STA]-x(3)-[LIVMFY] [D and R are the active site residues]-

[1] McAlister-Henn L. Trends Biochem. Sci. 13:178-181(1988).

[2] Gietl C. Biochim. Biophys. Acta 1100:217-234(1992).

10 [3] Birktoft J.J., Rhodes G., Banaszak L.J. Biochemistry 28:6065-6081(1989).

[4] Cendrin F., Chroboczek J., Zaccai G., Eisenberg H., Mevarech M. Biochemistry 32:4308-4313(1993).

5 334. Legume lectins signatures

Leguminous plants synthesize sugar-binding proteins which are called legume lectins [1,2].

These lectins are generally found in the seeds. The exact function of legume lectins is not known but they may be involved in the attachment of nitrogen-fixing bacteria to legumes and in the protection against pathogens. Legume lectins bind calcium and manganese (or other transition metals). Legume lectins are synthesized as precursor proteins of about 230 to 260 amino acid residues. Some legume lectins are proteolytically processed to produce two chains: beta (which corresponds to the N-terminal) and alpha (C-terminal). The lectin concanavalin A (conA) from jack bean is exceptional in that the two chains are transposed and ligated (by formation of a new peptide bond). The N-terminus of mature conA thus
20 corresponds to that of the alpha chain and the C-terminus to the beta chain. Two signature patterns specific to legume lectins have been developed: the first is located in the C-terminal section of the beta chain and contains a conserved aspartic acid residue important for the binding of calcium and manganese; the second one is located in the N-terminal of the alpha chain.

30

Consensus pattern: [LIV]-[STAG]-V-[DEQV]-[FLI]-D-[ST] [D binds manganese and calcium]-

Consensus pattern: [LIV]-x-[EDQ]-[FYWKR]-V-x-[LIVF]-G-[LF]-[ST]-

[1] Sharon N., Lis H. FASEB J. 4:3198-320(1990).

[2] Lis H., Sharon N. Annu. Rev. Biochem. 55:33-37(1986).

5

335. CoA-ligases (ligases- CoA)

This family includes the CoA ligases Succinyl-CoA synthetase alpha: and beta chains, malate CoA ligase and ATP-citrate lyase. Some members of the family utilise ATP others use GTP.

10 [1] Wolodko WT, Fraser ME, James MN, Bridger WA, J Biol Chem 1994;269:10883-10890.

336. linker histone H1 and H5 family

5 Linker histone H1 is an essential component of chromatin structure. H1 links nucleosomes into higher order structures Histone H1 is replaced by histone H5 in some cell types.

[1] Ramakrishnan V, Finch JT, Graziano V, Lee PL, Sweet RM, Nature 1993;362:219-223.

20

337. Lipocalin signature (lip1)

Proteins which transport small hydrophobic molecules such as steroids, bilins, retinoids, and lipids share limited regions of sequence homology and a common tertiary structure architecture [1 to 5]. This is an eight stranded antiparallel beta-barrel with a repeated + 1 topology enclosing a internal ligand binding site [1,3]. The name 'lipocalin' has been proposed [5] for this protein family. Proteins known to belong to this family are listed below (references are only provided for recently determined sequences). - Alpha-1-microglobulin (protein HC), which seems to bind porphyrin. - Alpha-1-acid glycoprotein (orosomucoid), which can bind a remarkable array of natural and synthetic compounds [6]. - Aphrodisin which, in hamsters, functions as an aphrodisiac pheromone. - Apolipoprotein D, which probably binds heme-related compounds. - Beta-lactoglobulin, a milk protein whose physiological function appears to bind retinol. - Complement component C8 gamma chain,

25

30

which seems to bind retinol [7]. - Crustacyanin [8], a protein from lobster carapace, which binds astaxanthin, a carotenoid. - Epididymal-retinoic acid binding protein (E-RABP) [9] involved in sperm maturation. - Insectacyanin, a moth bilin-binding protein, and a related butterfly bilin-binding protein (BBP). - Late Lactation protein (LALP), a milk protein from tammar wallaby [10]. - Neutrophil gelatinase-associated lipocalin (NGAL) (p25) (SV-40 induced 24p3 protein) [11]. - Odorant-binding protein (OBP), which binds odorants. - Plasma retinol-binding proteins (PRBP). - Human pregnancy-associated endometrial alpha-2 globulin. - Probasin (PB), a rat prostatic protein. - Prostaglandin D synthase (EC 5.3.99.2) (GSH-independent PGD synthetase), a lipocalin with enzymatic activity [12]. - Purpurin, a retinal protein which binds retinol and heparin. - Quiescence specific protein p20K from chicken (embryo CH21 protein). - Rodent urinary proteins (alpha-2-microglobulin), which may bind pheromones. - VNSP 1 and 2, putative pheromone transport proteins from mouse vomeronasal organ [13]. - Von Ebner's gland protein (VEGP) [14] (also called tear lipocalin), a mammalian protein which may be involved in taste recognition. - A frog olfactory protein, which may transport odorants. - A protein found in the cerebrospinal fluid of the toad Bufo Marinus with a supposed function similar to transthyretin in transport across the blood brain barrier [15]. - Lizard's epididymal secretory protein IV (LESP IV), which could transport small hydrophobic molecules into the epididymal fluid during sperm maturation [16]. - Prokaryotic outer-membrane protein ble [17]. The sequences of most members of the family, the core or kernel lipocalins, are characterized by three short conserved stretches of residues [3,18]. Others, the outlier lipocalin group, share only one or two of these [3,18]. A signature pattern was built around the first, common to all outlier and kernel lipocalins, which occurs near the start of the first beta-strand.

Consensus pattern: [DENG]-x-[DENQGSTARK]-x(0,2)-[DENQARK]-[LIVFY]-{CP}-G-{C}-W-[FYWLRH]-x-[LIVMTA]-

Note: it is suggested, on the basis of similarities of structure, function, and sequence, that this family forms an overall superfamily, called the calycins, with the avidin/streptavidin <PDOC00499> and the cytosolic fatty-acid binding proteins <PDOC00188> families [3,19]

[1] Cowan S.W., Newcomer M.E., Jones T.A. Proteins 8:44-61(1990).

[2] Igaraishi M., Nagata A., Toh H., Urade H., Hayaishi N. Proc. Natl. Acad. Sci. U.S.A. 89:5376-5380(1992).

- [3] Flower D.R., North A.C.T., Attwood T.K. Protein Sci. 2:753-761(1993).
- [4] Godovac-Zimmermann J. Trends Biochem. Sci. 13:64-66(1988).
- [5] Pervaiz S., Brew K. FASEB J. 1:209-214(1987).
- [6] Kremer J.M.H., Wilting J., Janssen L.H.M. Pharmacol. Rev. 40:1-47(1989).
- 5 [7] Haefliger J.-A., Peitsch M.C., Jenne D., Tschopp J. Mol. Immunol. 28:123-131(1991).
- [8] Keen J.N., Caceres I., Eliopoulos E.E., Zagalsky P.F., Findlay J.B.C. Eur. J. Biochem. 197:407-417(1991).
- [9] Newcomer M.E. Structure 1:7-18(1993).
- [10] Collet C., Joseph R. Biochim. Biophys. Acta 1167:219-222(1993).
- 10 [11] Kjeldsen L., Johnsen A.H., Sengelov H., Borregaard N. J. Biol. Chem. 268:10425-10432(1993).
- [12] Peitsch M.C., Boguski M.S. Trends Biochem. Sci. 16:363-363(1991).
- [13] Miyawaki A., Matsushita Y.R., Ryo Y., Mikoshiba T. EMBO J. 13:5835-5842(1994).
- [14] Kock K., Ahlers C., Schmale H. Eur. J. Biochem. 221:905-916(1994).
- 15 [15] Achen M.G., Harms P.J., Thomas T., Richardson S.J., Wettenhall R.E.H., Schreiber G. J. Biol. Chem. 267:23170-23174(1992).
- [16] Morel L., Dufarre J.-P., Depeiges A. J. Biol. Chem. 268:10274-10281(1993).
- [17] Bishop R.E., Penfold S.S., Frost L.S., Holtje J.V., Weiner J.H. J. Biol. Chem. 270:23097-23103(1995).
- 20 [18] Flower D.R., North A.C.T., Attwood T.K. Biochem. Biophys. Res. Commun. 180:69-74(1991).
- [19] Flower D.R. FEBS Lett. 333:99-102(1993).

Cytosolic fatty-acid binding proteins signature (lip2)

25 A number of low molecular weight proteins which bind fatty acids and other organic anions are present in the cytosol [1,2]. Most of them are structurally related and have probably diverged from a common ancestor. This structure is a ten stranded antiparallel beta-barrel, albeit with a wide discontinuity between the fourth and fifth strands, with a repeated + 1 topology enclosing an internal ligand binding site [2,7]. Proteins known to belong to this

30 family include: - Six, tissue-specific, types of fatty acid binding proteins (FABPs) found in liver, intestine, heart, epidermal, adipocyte, brain/retina. Heart FABP is also known as mammary-derived growth inhibitor (MDGI), a protein that reversibly inhibits proliferation of mammary carcinoma cells. Epidermal FABP is also known as psoriasis-associated FABP [3].

- Insect muscle fatty acid-binding proteins. - Testis lipid binding protein (TLBP). - Cellular retinol-binding proteins I and II (CRBP). - Cellular retinoic acid-binding protein (CRABP). - Gastrotropin, an ileal protein which stimulates gastric acid and pepsinogen secretion. It seems that gastrotropin binds to bile salts and bilirubins. - Fatty acid binding proteins MFB1 and MFB2 from the midgut of the insect *Manduca sexta* [4]. In addition to the above cytosolic proteins, this family also includes: - Myelin P2 protein, which may be a lipid transport protein in Schwann cells. P2 is associated with the lipid bilayer of myelin. - *Schistosoma mansoni* protein Sm14 [5] which seems to be involved in the transport of fatty acids. - *Ascaris suum* p18 a secreted protein that may play a role in sequestering potentially toxic fatty acids and their peroxidation products or that may be involved in the maintenance of the impermeable lipid layer of the eggshell. - Hypothetical fatty acid-binding proteins F40F4.2, F40F4.3, F40F4.4 and ZK742.5 from *Caenorhabditis elegans*. As a signature pattern for these proteins a segment from the N-terminal extremity was used.

Consensus pattern: [GSAIVK]-x-[FYW]-x-[LIVMF]-x(4)-[NHG]-[FY]-[DE]-x- [LIVMFY]-[LIVM]-x(2)-[LIVMAKR]-

Note: it is suggested, on the basis of similarities of structure, function, and sequence, that this family forms an overall superfamily, called the calycins, with the lipocalin <PDOC00187> and avidin/streptavidin <PDOC00499> families [6,7].

[1] Bernier I., Jolles P. *Biochimie* 69:1127-1152(1987).

[2] Veerkamp J.H., Peeters R.A., Maatman R.G.H.J. *Biochim. Biophys. Acta* 1081:1-24(1991).

[3] Siegenthaler G., Hotz R., Chatellard-Gruaz D., Didierjean L., Hellman U., Saurat J.-H. *Biochem. J.* 302:363-371(1994).

[4] Smith A.F., Tsuchida K., Hanneman E., Suzuki T.C., Wells M.A. *J. Biol. Chem.* 267:380-384(1992).

[5] Moser D., Tendler M., Griffiths G., Klinkert M.-Q. *J. Biol. Chem.* 266:8447-8454(1991).

[6] Flower D.R., North A.C.T, Attwood T.K. *Protein Sci.* 2:753-761(1993).

[7] Flower D.R. *FEBS Lett.* 333:99-102(1993).

Lipoxygenases (EC 1.13.11.-) are a class of iron-containing dioxygenases which catalyzes the hydroperoxidation of lipids, containing a cis,cis-1,4-pentadiene structure. They are common in plants where they may be involved in a number of diverse aspects of plant physiology including growth and development, pest resistance, and senescence or responses to wounding [1]. In mammals a number of lipoxygenases isozymes are involved in the metabolism of prostaglandins and leukotrienes [2]. Sequence data is available for the following lipoxygenases: - Plant lipoxygenases (EC 1.13.11.12). Plants express a variety of cytosolic isozymes as well as what seems [3] to be a chloroplast isozyme. - Mammalian arachidonate 5-lipoxygenase (EC 1.13.11.34). - Mammalian arachidonate 12-lipoxygenase (EC 1.13.11.31). - Mammalian erythroid cell-specific 15-lipoxygenase (EC 1.13.11.33). The iron atom in lipoxygenases is bound by four ligands, three of which are histidine residues [4]. Six histidines are conserved in all lipoxygenase sequences, five of them are found clustered in a stretch of 40 amino acids. This region contains two of the three zinc-ligands; the other histidines have been shown [5] to be important for the activity of lipoxygenases. As signatures for this family of enzymes two patterns in the region of the histidine cluster were selected. The first pattern contains the first three conserved histidines and the second pattern includes the fourth and the fifth.

Consensus pattern: H-[EQ]-x(3)-H-x-[LM]-[NQRC]-[GST]-H-[LIVMSTAC](3)-E [The second and third H's bind iron]-

Consensus pattern: [LIVMA]-H-P-[LIVM]-x-[KRQ]-[LIVMF](2)-x-[AP]-H-

[1] Vick B.A., Zimmerman D.C. (In) Biochemistry of plants: A comprehensive treatise, Stumpf P.K., Ed., Vol. 9, pp.53-90, Academic Press, New-York, (1987).

[2] Needleman P., Turk J., Jakschik B.A., Morrison A.R., Lefkowitz J.B. Annu. Rev. Biochem. 55:69-102(1986).

[3] Peng Y.L., Shirano Y., Ohta H., Hibino T., Tanaka K., Shibata D. J. Biol. Chem. 269:3755-3761(1994).

[4] Boyington J.C., Gaffney B.J., Amzel L.M. Science 260:1482-1486(1993).

[5] Steczko J., Donoho G.P., Clemens J.C., Dixon J.E., Axelrod B. Biochemistry 31:4053-4057(1992).

339. Fumarate lyases signature (lyase_1)

A number of enzymes, belonging to the lyase class, for which fumarate is a substrate have been shown [1,2] to share a short conserved sequence around a methionine which is probably involved in the catalytic activity of this type of enzymes. These enzymes are: - Fumarase (EC 4.2.1.2) (fumarate hydratase), which catalyzes the reversible hydration of fumarate to L-malate. There seem to be 2 classes of fumarases: class I are thermolabile dimeric enzymes (as for example: *Escherichia coli* fumC); class II enzymes are thermostable and tetrameric and are found in prokaryotes (as for example: *Escherichia coli* fumA and fumB) as well as in eukaryotes. The sequence of the two classes of fumarases are not closely related. - Aspartate ammonia-lyase (EC 4.3.1.1) (aspartase), which catalyzes the reversible conversion of aspartate to fumarate and ammonia. This reaction is analogous to that catalyzed by fumarase, except that ammonia rather than water is involved in the trans-elimination reaction. -

Argininosuccinase (EC 4.3.2.1) (argininosuccinate lyase), which catalyzes the formation of arginine and fumarate from argininosuccinate, the last step in the biosynthesis of arginine. -

Adenylosuccinase (EC 4.3.2.2) (adenylosuccinate lyase) [3], which catalyzes the eighth step in the de novo biosynthesis of purines, the formation of 5'-phosphoribosyl-5-amino-4-imidazolecarboxamide and fumarate from 1-(5-phosphoribosyl)-4-(N-succinyl-carboxamide). That enzyme can also catalyze the formation of fumarate and AMP from adenylosuccinate. -

Pseudomonas putida 3-carboxy-cis,cis-muconate cycloisomerase (EC 5.5.1.2) (3-carboxymuconate lactonizing enzyme) (gene *pcaB*) [4], an enzyme involved in aromatic acids catabolism

Consensus pattern: G-S-x(2)-M-x(2)-K-x-N-

[1] Woods S.A., Schwartzbach S.D., Guest J.R. Biochim. Biophys. Acta 954:14-26(1988).

[2] Woods S.A., Miles J.S., Guest J.R. FEMS Microbiol. Lett. 51:181-186(1988).

[3] Zalkin H., Dixon J.E. Prog. Nucleic Acid Res. Mol. Biol. 42:259-287(1992).

[4] Williams S.E., Woolridge E.M., Ransom S.C., Landro J.A., Babbitt P.C., Kozarich J.W. Biochemistry 31:9768-9776(1992).

340. MCM family signature and profile

Proteins shown to be required for the initiation of eukaryotic DNA replication share a highly conserved domain of about 210 amino-acid residues [1,2,3]. The latter shows some similarities [4] with that of various other families of DNA-dependent ATPases. Eukaryotes seem to possess a family of six proteins that contain this domain. They were first identified in yeast where most of them have a direct role in the initiation of chromosomal DNA replication by interacting directly with autonomously replicating sequences (ARS). They were thus called 'minichromosome maintenance proteins' with gene symbols prefixed by MCM. These six proteins are: - MCM2, also known as cdc19 (in *S.pombe*) [E1]. - MCM3, also known as DNA polymerase alpha holoenzyme-associated protein P1, RLF beta subunit or ROA. - MCM4, also known as CDC54, cdc21 (in *S.pombe*) or dpa (in *Drosophila*). - MCM5, also known as CDC46 or nda4 (in *S.pombe*). - MCM6, also known as mis5 (in *S.pombe*). - MCM7, also known as CDC47 or Prolifera (in *A.thaliana*). This family is also present in archebacteria. In *Methanococcus jannaschii* there are four members: MJ0363, MJ0961, MJ1489 and MJECL13. The presence of a putative ATP-binding domain implies that these proteins maybe involved in an ATP-consuming step in the initiation of DNA replication in eukaryotes. As a signature pattern, a perfectly conserved region was selected that represents a special version of the B motif found in ATP-binding proteins.

Consensus pattern: G-[IVT]-[LVAC](2)-[IVT]-D-[DE]-[FL]-[DNST]

[1] Coxon A., Maundrell K., Kearsey S.E. *Nucleic Acids Res.* 20:5571-5577(1992).

[2] Hu B., Burkhart R., Schulte D., Musahl C., Knippers R. *Nucleic Acids Res.* 21:5289-5293(1993).

[3] Tye B.-K. *Trends Cell Biol.* 4:160-166(1994).

[4] Koonin E.V. *Nucleic Acids Res.* 21:2541-2547(1993).

341. Macrophage migration inhibitory factor family signature (MIF)

A protein called macrophage migration inhibitory factor (MIF) [1] seems to exert an important role in host inflammatory responses. It play a pivotal role in the host response to endotoxic shock and appears to serve as a pituitary "stress" hormone that regulates systemic inflammatory responses. MIF is a secreted protein of 115 residues which is not processed from a larger precursor. D-dopachrome tautomerase [2] is a mammalian cytoplasmic enzyme

involved in melanin biosynthesis and that tautomerizes D-dopachrome with concomitant decarboxylation to give 5,6-dihydroxyindole (DHI). It is a protein of 117 residues highly related to MIF. It must be noted that MIF binds glutathione and has been said to be related to glutathione S-transferases. This assertion has been later disproved [3]. As a signature pattern for these proteins, a conserved region was selected located in the central section.

Consensus pattern: [DE]-P-C-A-x(3)-[LIVM]-x-S-I-G-x-[LIVM]-G-

[1] Bucala R. Immunol. Lett. 43:23-26(1994).

[2] Odh G., Hindemith A., Rosengren A.-M., Rosengren E., Rorsman H. Biochem. Biophys. Res. Commun. 197:619-624(1993).

[3] Pearson W.R. Protein Sci. 3:525-527(1994).

342. MIP family signature

Recently the sequence of a number of different proteins, that all seem to be transmembrane channel proteins, has been found to be highly related [1 to 4]. These proteins are listed below.

- Mammalian major intrinsic protein (MIP). MIP is the major component of lens fiber gap junctions. Gap junctions mediate direct exchange of ions and small molecule from one cell to another.
- Mammalian aquaporins [5]. These proteins form water-specific channels that provide the plasma membranes of red cells and kidney proximal and collecting tubules with high permeability to water, thereby permitting water to move in the direction of an osmotic gradient.
- Soybean nodulin-26, a major component of the peribacteroid membrane induced during nodulation in legume roots after Rhizobium infection.
- Plants tonoplast intrinsic proteins (TIP). There are various isoforms of TIP: alpha (seed), gamma, Rt (root), and Wsi (water-stress induced). These proteins may allow the diffusion of water, amino acids and/or peptides from the tonoplast interior to the cytoplasm.
- Bacterial glycerol facilitator protein (gene glpF), which facilitates the movement of glycerol across the cytoplasmic membrane.
- Salmonella typhimurium propanediol diffusion facilitator (gene pduF).
- Yeast FPS1, a glycerol uptake/efflux facilitator protein.
- Drosophila neurogenic protein 'big brain' (bib). This protein may mediate intercellular communication; it may functions by allowing the transport of certain molecules(s) and thereby sending a signal for an exodermal cell to become an epidermoblast instead of a neuroblast.
- Yeast hypothetical protein YFL054c.

hypothetical protein from the pepX region of lactococcus lactis. The MIP family proteins seem to contain six transmembrane segments. Computer analysis shows that these protein probably arose by a tandem, intragenic duplication event from an ancestral protein that contained three transmembrane segments. As a signature pattern a well conserved region was selected which is located in a probable cytoplasmic loop between the second and third transmembrane regions.

Consensus pattern: [HNQA]-x-N-P-[STA]-[LIVMF]-[ST]-[LIVMF]-[GSTAFY]-

- [1] Reizer J., Reizer A., Saier M.H. Jr. CRC Crit. Rev. Biochem. 28:235-257(1993).
- [2] Baker M.E., Saier M.H. Jr. Cell 60:185-186(1990).
- [3] Pao G.M., Wu L.-F., Johnson K.D., Hoeft H., Chrispeels M.J., Sweet G., Sandal N.N., Saier M.H. Jr. Mol. Microbiol. 5:33-37(1991).
- [4] Wistow G.J., Pisano M.M., Chepelinsky A.B. Trends Biochem. Sci. 16:170-171(1991).
- [5] Chrispeels M.J., Agre P. Trends Biochem. Sci. 19:421-425(1994).

343. Mandelate racemase / muconate lactonizing enzyme family signatures

Mandelate racemase (EC 5.1.2.2) (MR) and muconate lactonizing enzyme (EC 5.5.1.1) (MLE) are two bacterial enzymes involved in aromatic acid catabolism. They catalyze mechanistically distinct reactions yet they are related at the level of their primary, quaternary (homooctamer) and tertiary structures [1,2]. A number of other proteins also seem to be evolutionary related to these two enzymes. These are: - The various plasmid-encoded chloromuconate cycloisomerases (EC 5.5.1.7). - Escherichia coli protein rspA [3], rspA seems to be involved in the degradation of homoserine lactone (HSL) or of one of its metabolite. - Escherichia coli hypothetical protein ycjG. - Escherichia coli hypothetical protein yidU. - A hypothetical protein from Streptomyces ambofaciens [4]. Two signature patterns have been developed for these enzymes; both contain conserved acidic residues. The second pattern contains an aspartate and a glutamate which are ligands for either a magnesium ion (in MR) or a manganese ion (in MLE).

Consensus pattern: A-x-[SAGCN]-[SAG]-[LIVM]-[DEQ]-x-A-[LA]-x-[DE]-[LIA]-x-[GA]-[KRQ]-x(4)-[PSA]-[LIV]-x(2)-L-[LIVMF]-G-

Consensus pattern: [LIVF]-x(2)-D-x-[NH]-x(7)-[ACL]-x(6)-[LIVMF]-x(7)-[LIVM]- E-
[DENQ]-P [D and E bind a divalent metal ion]-

[1] Neidhart D.J., Kenyon G.L., Gerlt J.A., Petsko G.A. Nature 347:692-694(1990).

5 [2] Petsko G.A., Kenyon G.L., Gerlt J.A., Ringe D., Kozarich J.W. Trends Biochem. Sci.
18:372-376(1993).

[3] Huisman G.W., Kolter R. Science 265:537-539(1994).

[4] Schneider D., Aigle B., Leblond P., Simonet J.M., Decaris B. J. Gen. Microbiol.
139:2559-2567(1993).

10

344. Merozoite Surface Antigen 2 (MSA-2) family

Thomas AW, Carr DA, Carter JM, Lyon JA, Mol Biochem Parasitol 1990;43:211-
220.

345. MSP (Major sperm protein) domain.

Major sperm proteins are involved in sperm motility. These proteins oligomerise to
form filaments. Partial matches to this domain are also found in other non MSP proteins.

These include Swiss:P40075 and Swiss:P34593.

[1] Bullock TL, Roberts TM, Stewart M, J Mol Biol 1996;263:284-296. [2] King KL,
Stewart M, Roberts TM, Seavy M, J Cell Sci 1992;101:847-857.

25 346. (Matrix) Viral matrix protein. Found in Morbillivirus and paramyxovirus, pneumovirus.
Number of members: 105

347. O-methyltransferase (methyltransf)

30 This family includes a range of O-methyltransferases. These enzymes utilise S-
adenosyl methionine.

[1] Keller NP, Dischinger HC, Bhatnagar D, Cleveland TE, Ullah AH, Appl Environ
Microbiol 1993;59:479-484.

348. Magnesium chelatase, subunit ChII

Magnesium-chelatase is a three-component enzyme that catalyses the insertion of Mg²⁺ into protoporphyrin IX. This is the first unique step in the synthesis of (bacterio)chlorophyll. Due to this, it is thought that Mg-chelatase has an important role in channeling inter- mediates into the (bacterio)chlorophyll branch in response to conditions suitable for photosynthetic growth. ChII and BchD have molecular weight between 38-42 kDa.

[1] Walker CJ, Willows RD, Biochem J 1997;327:321-333. [2] Petersen BL, Jensen PE, Gibson LC, Stummann BM, Hunter CN, Henningsen KW, J Bacteriol 1998;180:699-704.

349. Plasmid recombination enzyme (Mob_Pre)

With some plasmids, recombination can occur in a site specific manner that is independent of RecA. In such cases, the recombination event requires another protein called Pre. Pre is a plasmid recombination enzyme. This protein is: also known as Mob (conjugative mobilization).

[1] Priebe SD, Lacks SA, J Bacteriol 1989;171:4778-4784.

350. Monooxygenase

This family includes diverse enzymes that utilise FAD.

[1] Gatti DL, Palfey BA, Lah MS, Entsch B, Massey V, Ballou DP, Ludwig ML, Science 1994;266:110-114.

351. Mov34 family

Members of this family are found in proteasome regulatory subunits, eukaryotic initiation factor 3 (eIF3) subunits and regulators of transcription factors.

[1] Aravind L, Ponting CP, Protein Sci 1998;7:1250-1254. [2] Hershey JW, Asano K, Naranda T, Vornlocher HP, Hanachi P, Merrick WC, Biochimie 1996;78:903-907.

352. Myc amino-terminal region (Myc_N_term)

The myc family belongs to the basic helix-loop-helix leucine zipper class of transcription factors, see HLH. Myc forms a heterodimer with Max, and this complex regulates cell growth through direct activation of genes involved in cell replication [2].

[1] Facchini LM, Penn LZ, FASEB J 1998;12:633-651. [2] Grandori C, Eisenman RN, Trends Biochem Sci 1997;22:177-181.

353. (Metallothio_2) Metallothionein. Members of this family are metallothioneins. These proteins are cysteine rich proteins that bind to heavy metals. Members of this family appear to be closest to Class II metallothioneins, seed metalthio. Number of members: 55

[1] Medline: 98267202. Characterization of gene repertoires at mature stage of citrus fruits through random sequencing and analysis of redundant metallothionein-like genes expressed during fruit development. Moriguchi T, Kita M, Hisada S, Endo-Inagaki T, Omura M; Gene 1998;211:221-227.

354. MAGE family

The MAGE (melanoma antigen-encoding gene) family are expressed in a wide variety of tumors but not in normal cells, with the exception of the male germ cells, placenta, and, possibly, cells of the developing embryo. The cellular function of this family is unknown.

[1] McCurdy DK, Tai LQ, Nguyen J, Wang Z, Yang HM, Udar N, Naiem F, Concannon P, Gatti RA; Mol Genet Metab 1998;63:3-13.

355. Malic enzymes signature. Malic enzymes, or malate oxidoreductases, catalyze the oxidative decarboxylation of malate into pyruvate important for a wide range of metabolic

pathways. There are three related forms of malic enzyme [1,2,3]: - NAD-dependent malic enzyme (EC 1.1.1.38), which uses preferentially NAD and has the ability to decarboxylate oxaloacetate (OAA). It is found in bacteria and insects. - NAD-dependent malic enzyme (EC 1.1.1.39), which uses preferentially NAD and is unable to decarboxylate OAA. It is found in the mitochondrial matrix of plants and is a heterodimer of highly related subunits. - NADP-dependent malic enzyme (EC 1.1.1.40), which has a preference for NADP and has the ability to decarboxylate OAA. This form has been found in fungi, animals and plants. In mammals, there are two isozymes: one, mitochondrial and the other, cytosolic. Plants also have two isozymes: chloroplastic and cytosolic. There are two other proteins which are closely structurally related to malicenzymes: - Escherichia coli protein sfcA, whose function is not yet known but which could be an NAD or NADP-dependent malic enzyme. - Yeast hypothetical protein YKL029c, a probable malic enzyme. There are three well conserved regions in the enzyme sequences. Two of them seem to be involved in binding NAD or NADP. The significance of the third one, located in the central part of the enzymes, is not yet known. This region has been developed as a signature pattern for these enzymes.

Consensus pattern: F-x-[DV]-D-x(2)-G-T-[GSA]-x-[IV]-x-[LIVMA]-[GAST](2)-[LIVMF](2)-

[1] Artus N.N., Edwards G.E. FEBS Lett. 182:225-233(1985).[2] Loeber G., Infante A.A., Maurer-Fogy I., Krystek E., Dworkin M.B. J. Biol. Chem. 266:3016-3021(1991). [3] Long J.J., Wang J.-L., Berry J.O. J. Biol. Chem. 269:2827-2833(1994).

356. (matrixin)

Matrixins cysteine switch (aka peptidase_M10)

Mammalian extracellular matrix metalloproteinases (EC 3.4.24.-), also known as matrixins [1] (see <PDOC00129>), are zinc-dependent enzymes. They are secreted by cells in an inactive form (zymogen) that differs from the mature enzyme by the presence of an N-terminal propeptide. A highly conserved octapeptide is found two residues downstream of the C-terminal end of the propeptide. This region has been shown to be involved in autoinhibition of matrixins [2,3]; a cysteine within the octapeptide chelates the active site

zinc ion, thus inhibiting the enzyme. This region has been called the 'cysteine switch' or 'autoinhibitor region'.

A cysteine switch has been found in the following zinc proteases:

- MMP-1 (EC 3.4.24.7) (interstitial collagenase).
- MMP-2 (EC 3.4.24.24) (72 Kd gelatinase).
- MMP-3 (EC 3.4.24.17) (stromelysin-1).
- MMP-7 (EC 3.4.24.23) (matrilysin).
- MMP-8 (EC 3.4.24.34) (neutrophil collagenase).
- MMP-9 (EC 3.4.24.35) (92 Kd gelatinase).
- MMP-10 (EC 3.4.24.22) (stromelysin-2).
- MMP-11 (EC 3.4.24.-) (stromelysin-3).
- MMP-12 (EC 3.4.24.65) (macrophage metalloelastase).
- MMP-13 (EC 3.4.24.-) (collagenase 3).
- MMP-14 (EC 3.4.24.-) (membrane-type matrix metalloproteinase 1).
- MMP-15 (EC 3.4.24.-) (membrane-type matrix metalloproteinase 2).
- MMP-16 (EC 3.4.24.-) (membrane-type matrix metalloproteinase 3).
- Sea urchin hatching enzyme (EC 3.4.24.12) (envelysin) [4].
- Chlamydomonas reinhardtii gamete lytic enzyme (GLE) [5].

Consensus pattern P-R-C-[GN]-x-P-[DR]-[LIVSAPKQ] [C chelates the zinc ion] Sequences known to belong to this class detected by the pattern ALL, except for cat MMP-7 and mouse MMP-11.

[1] Woessner J. Jr. FASEB J. 5:2145-2154(1991).

□

[2] Sanchez-Lopez R., Nicholson R., Gesnel M.C., Matrisian L.M., Breathnach R. J. Biol. Chem. 263:11892-11899(1988).

[3] Park A.J., Matrisian L.M., Kells A.F., Pearson R., Yuan Z., Navre M. J. Biol. Chem. 266:1584-1590(1991).

[4] Lepage T., Gache C. EMBO J. 9:3003-3012(1990).

[5] Kinoshita T., Fukuzawa H., Shimada T., Saito T., Matsuda Y. Proc. Natl. Acad. Sci. U.S.A. 89:4693-4697(1992).

5 357. Vertebrate metallothioneins signature (metalthio)

Metallothioneins (MT) [1,2,3] are small proteins which bind heavy metals such as zinc, copper, cadmium, nickel, etc., through clusters of thiolate bonds. MT's occur throughout the animal kingdom and are also found in higher plants, fungi and some prokaryotes. On the basis of structural relationships MT's have been subdivided into three classes. Class I includes
 10 mammalian MT's as well as MT's from crustacean and molluscs, but with clearly related primary structure. Class II groups together MT's from various species such as sea urchins, fungi, insects and cyanobacteria which display none or only very distant correspondence to class I MT's. Class III MT's are atypical polypeptides containing gamma-glutamylcysteinyl units. Vertebrate class I MT's are proteins of 60 to 68 amino acid residues, 20 of these
 15 residues are cysteines that bind to 7 bivalent metal ions. As a signature pattern a region that spans 19 residues and which contains seven of the metal-binding cysteines was chosen, this region is located in the N-terminal section of class-I MT's.

Consensus pattern: C-x-C-[GSTAP]-x(2)-C-x-C-x(2)-C-x-C-x(2)-C-x-K-

[1] Hamer D.H. Annu. Rev. Biochem. 55:913-951(1986).

[2] Kagi J.H.R., Schaffer A. Biochemistry 27:8509-8515(1988).

[3] Binz P.-A. Thesis, 1996, University of Zurich.

25

358. Mitochondrial energy transfer proteins signature (mito_carr)

Different types of substrate carrier proteins involved in energy transfer are found in the inner mitochondrial membrane [1 to 5]. These are: - The ADP,ATP carrier protein (AAC) (ADP/ATP translocase) which exports ATP into the cytosol and imports ADP into the
 30 mitochondrial matrix. The sequence of AAC has been obtained from various mammalian, plant and fungal species. - The 2-oxoglutarate/malate carrier protein (OGCP), which exports 2-oxoglutarate into the cytosol and imports malate or other dicarboxylic acids into the mitochondrial matrix. This protein plays an important role in several metabolic processes

such as the malate/aspartate and the oxoglutarate/isocitrate shuttles. - The phosphate carrier protein, which transports phosphate groups from the cytosol into the mitochondrial matrix. - The brown fat uncoupling protein (UCP) which dissipates oxidative energy into heat by transporting protons from the cytosol into the mitochondrial matrix. - The tricarboxylate transport protein (or citrate transport protein) which is involved in citrate-H⁺/malate exchange. It is important for the bioenergetics of hepatic cells as it provides a carbon source for fatty acid and sterol biosyntheses, and NAD for the glycolytic pathway. - The Grave's disease carrier protein (GDC), a protein of unknown function recognized by IgG in patients with active Grave's disease. - Yeast mitochondrial proteins MRS3 and MRS4. The exact function of these proteins is not known. They suppress a mitochondrial splice defect in the first intron of the COB gene and may act as carriers, exerting their suppressor activity by modulating solute concentrations in the mitochondrion. - Yeast mitochondrial FAD carrier protein (gene FLX1). - Yeast protein ACR1 [6], which seems essential for acetyl-CoA synthetase activity. - Yeast protein PET8. - Yeast protein PMT. - Yeast protein RIM2. - Yeast protein YHM1/SHM1. - Yeast protein YMC1. - Yeast protein YMC2. - Yeast hypothetical proteins YBR291c, YEL006w, YER053c, YFR045w, YHR002w, and YIL006w. - *Caenorhabditis elegans* hypothetical protein K11H3.3. Two other proteins have been found to belong to this family, yet are not localized in the mitochondrial inner membrane: - Maize amyloplast Brittle-1 protein. This protein, found in the endosperm of kernels, could play a role in amyloplast membrane transport. - *Candida boidinii* peroxisomal membrane protein PMP47 [7]. PMP47 is an integral membrane protein of the peroxisome and it may play a role as a transporter. These proteins all seem to be evolutionary related. Structurally, they consist of three tandem repeats of a domain of approximately one hundred residues. Each of these domains contains two transmembrane regions. As a signature pattern, one of the most conserved regions in the repeated domain was selected, located just after the first transmembrane region.

Consensus pattern: P-x-[DE]-x-[LIVAT]-[RK]-x-[LRH]-[LIVMFY]-[QGAIVM]-

- [1] Klingenberg M. Trends Biochem. Sci. 15:108-112(1990).
- [2] Walker J.E. Curr. Opin. Struct. Biol. 2:519-526(1992).
- [3] Kuan J., Saier M.H. Jr. CRC Crit. Rev. Biochem. 28:209-233(1993).
- [4] Kuan J., Saier M.H. Jr. Res. Microbiol. 144:671-672(1993).

[5] Nelson D.R., Lawson J.E., Klingenberg M., Douglas M.G. J. Mol. Biol. 230:1159-1170(1993).

[6] Palmieri F. FEBS Lett. 346:48-54(1994).

[7] Jank B., Habermann B., Schweyen R.J., Link T.A. Trends Biochem. Sci. 18:427-428(1993).

359. Prokaryotic molybdopterin oxidoreductases signatures (molybdopterin)

A number of different prokaryotic oxidoreductases that require and bind amolybdopterin cofactor have been shown [1,2,3] to share a number of regions of sequence similarity. These enzymes are: - *Escherichia coli* respiratory nitrate reductase (EC 1.7.99.4). This enzyme complex allows the bacteria to use nitrate as an electron acceptor during anaerobic growth. The enzyme is composed of three different chains: alpha, beta and gamma. The alpha chain (gene *narG*) is the molybdopterin-binding subunit. *Escherichia coli* encodes for a second, closely related, nitrate reductase complex which also contains a molybdopterin-binding alpha chain (gene *narZ*). - *Escherichia coli* anaerobic dimethyl sulfoxide reductase (DMSO reductase). DMSO reductase is the terminal reductase during anaerobic growth on various sulfoxide and N-oxide compounds. DMSO reductase is composed of three chains: A, B and C. The A chain (gene *dmsA*) binds molybdopterin. - *Escherichia coli* biotin sulfoxide reductases (genes *bisC* and *bisZ*). This enzyme reduces a spontaneous oxidation product of biotin, BDS, back to biotin. It may serve as a scavenger, allowing the cell to use biotin sulfoxide as a biotin source. - *Methanobacterium formicicum* formate dehydrogenase (EC 1.2.1.2). The alpha chain (gene *fdhA*) of this dimeric enzyme binds a molybdopterin cofactor. - *Escherichia coli* formate dehydrogenases -H (gene *fdhF*), -N (gene *fdnG*) and -O (gene *fdoG*). These enzymes are responsible for the oxidation of formate to carbon dioxide. In addition to molybdopterin, the alpha (catalytic) subunit also contains an active site, selenocysteine. - *Wolinella succinogenes* polysulfide reductase chain. This enzyme is a component of the phosphorylative electron transport system with polysulfide as the terminal acceptor. It is composed of three chains: A, B and C. The A chain (gene *psrA*) binds molybdopterin. - *Salmonella typhimurium* thiosulfate reductase (gene *phsA*). - *Escherichia coli* trimethylamine-N-oxide reductase (EC 1.6.6.9) (gene *torA*) [4]. - Nitrate reductase (EC 1.7.99.4) from *Klebsiella pneumoniae* (gene *nasA*), *Alcaligenes eutrophus*, *Escherichia coli*, *Rhodobacter sphaeroides*, *Thiosphaera pantotropha* (gene *napA*), and *Synechococcus* PCC

7942 (gene narB). These proteins range from 715 amino acids (fdhF) to 1246 amino acids (narZ) in size. Three signature patterns for these enzymes were derived. The first is based on a conserved region in the N-terminal section and contains two cysteine residues perhaps involved in binding the molybdopterin cofactor. It should be noted that this region is not present in bisC. The second pattern is derived from a conserved region located in the central part of these enzymes.

Consensus pattern: [STAN]-x-[CH]-x(2,3)-C-[STAG]-[GSTVMF]-x-C-x-[LIVMFYW]-x-[LIVMA]-x(3,4)-[DENQKHT]-

Consensus pattern: [STA]-x-[STAC](2)-x(2)-[STA]-D-[LIVMY](2)-L-P-x-[STAC](2)-x(2)-E-

Consensus pattern: A-x(3)-[GDT]-I-x-[DNQTK]-x-[DEA]-x-[LIVM]-x-[LIVMC]-x-[NS]-x(2)-[GS]-x(5)-A-x-[LIVM]-[ST]-

[1] Wootton J.C., Nicolson R.E., Cock J.M., Walters D.E., Burke J.F., Doyle W.A., Bray R.C. Biochim. Biophys. Acta 1057:157-185(1991).

[2] Bilous P.T., Cole S.T., Anderson W.F., Weiner J.H. Mol. Microbiol. 2:785-795(1988).

[3] Trieber C.A., Rothery R.A., Weiner J.H. J. Biol. Chem. 269:7103-7109(1994).

[4] Mejean V., Lobbi-Nivol C., Lepelletier M., Giordano G., Chippaux M., Pascal M.-C. Mol. Microbiol. 11:1169-1179(1994).

360. Bacterial mutT domain signature

The bacterial mutT protein is involved in the GO system [1] responsible for removing an oxidatively damaged form of guanine (8-hydroxyguanine or 7,8-dihydro-8-oxoguanine) from DNA and the nucleotide pool. 8-oxo-dGTP is inserted opposite to dA and dC residues of template DNA with almost equal efficiency thus leading to A.T to G.C transversions. MutT specifically degrades 8-oxo-dGTP to the monophosphate with the concomitant release of pyrophosphate. MutT is a small protein of about 12 to 15 Kd. It has been shown [2,3] that a region of about 40 amino acid residues, which is found in the N-terminal part of mutT, can also be found in a variety of other prokaryotic, viral, and eukaryotic proteins. These proteins are:

- *Streptomyces pneumoniae* mutX.

352

- A mutT homolog from plasmid pSAM2 of *Streptomyces ambofaciens*.
- *Bartonella bacilliformis* invasion protein A (gene *invA*).
- *Escherichia coli* dATP pyrophosphohydrolase.
- Protein D250 from African swine fever viruses.
- Proteins D9 and D10 from a variety of poxviruses.
- Mammalian 7,8-dihydro-8-oxoguanine triphosphatase (EC 3.1.6.-) [4].
- Mammalian diadenosine 5',5'''-P1,P4-tetraphosphate asymmetrical hydrolase (Ap4Aase) (EC 3.6.1.17) [5], which cleaves A-5'-PPPP-5'A to yield AMP and ATP.
- A protein encoded on the antisense RNA of the basic fibroblast growth factor gene in higher vertebrates.
- Yeast protein YSA1.
- *Escherichia coli* hypothetical protein *yfaO*.
- *Escherichia coli* hypothetical protein *ygdU* and HI0901, the corresponding *Haemophilus influenzae* protein.
- *Escherichia coli* hypothetical protein *yjaD* and HI0432, the corresponding *Haemophilus influenzae* protein.
- *Escherichia coli* hypothetical protein *yrfE*.
- *Bacillus subtilis* hypothetical protein *yqkG*.
- *Bacillus subtilis* hypothetical protein *yzgD*.
- Yeast hypothetical protein YGL067w.

It is proposed [2] that the conserved domain could be involved in the active center of a family of pyrophosphate-releasing NTPases. As a signature pattern the core region of the domain was selected; it contains four conserved glutamate residues.

Consensus pattern: G-x(5)-E-x(4)-[STAGC]-[LIVMAC]-x-R-E-[LIVMFT]-x-E-E-

[1] Michaels M.L., Miller J.H. J. Bacteriol. 174:6321-6325(1992).

[2] Koonin E.V. Nucleic Acids Res. 21:4847-4847(1993).

[3] Mejean V., Salles C., Bullions M.J., Bessman M.J., Claverys J.-P. Mol. Microbiol. 11:323-330(1994).

[4] Sakumi K., Furuichi M., Tsuzuki T., Kakuma T., Kawabata S., Maki H., Sekiguchi M. J. Biol. Chem. 268:23524-23530(1993).

[5] Thorne N.M.H., Hankin S., Wilkinson M.C., Nunez C., Barraclough R., McLennan A.G. Biochem. J. 311:717-721(1995).

361. Myb DNA-binding domain repeat signatures

The retroviral oncogene v-myb, and its cellular counterpart c-myb, encode nuclear DNA-binding proteins that specifically recognize the sequence YAAC(G/T)G [1]. The myb family also includes the following proteins: - Drosophila D-myb [2]. - Vertebrate myb-like proteins A-myb and B-myb [3]. - Maize C1 protein, a trans-acting factor which controls the expression of genes involved in anthocyanin biosynthesis. - Maize P protein [4], a trans-acting factor which regulates the biosynthetic pathway of a flavonoid-derived pigment in certain floral tissues. - Arabidopsis thaliana protein GL1 [5], required for the initiation of differentiation of leaf hair cells (trichomes). - A number of myb/c1-related proteins in maize and barley, whose roles are not yet known [4]. - Yeast BAS1 [7], a transcriptional activator for the HIS4 gene. - Yeast REB1 [8], which recognizes sites within both the enhancer and the promoter of rRNA transcription, as well as upstream of many genes transcribed by RNA polymerase II. - Fission yeast cdc5, a possible transcription factor whose activity is required for cell cycle progression and growth during G2. - Fission yeast myb1, which regulates telomere length and function. - Yeast hypothetical protein YMR213w. One of the most conserved regions in all of these proteins is a domain of 160 amino acids. It consists of three tandem repeats of 51 to 53 amino acids. In myb, this repeat region has been shown [9] to be involved in DNA-binding. The major part of the first repeat is missing in retroviral v-myb sequences and in plant myb-related proteins. Yeast REB1 differs from the other proteins in this family in having a single myb-like domain. As shown in the following schematic representation, two signature patterns for myb-like domains were developed; the first is located in the N-terminal section, the second spans the C-terminal extremity of the domain.

xxxxxxxxWxxxEDxxxxxxxxxxxxxxxxWxxIxxxxxxxxRxxxxxxxxWxxxx *****

*****! : Position of the patterns.

Consensus pattern: W-[ST]-x(2)-E-[DE]-x(2)-[LIV]-

Consensus pattern: W-x(2)-[LI]-[SAG]-x(4,5)-R-x(8)-[YW]-x(3)-[LIVM]-

Note: this pattern detects the three copies of the domain in myb, d-myb, A-myb and B-myb; the second of the two complete copies of plant myb-related proteins, and the last two copies of yeast BAS1

- [1] Biednkapp H., Borgmeyer U., Sippel A.E., Klempnauer K.-H. Nature 335:835-837(1988).
- [2] Peters C.W.B., Sippel A.E., Vingron M., Klempnauer K.-H. EMBO J. 6:3085-3090(1987).
- [3] Nomura N., Takahashi M., Matsui M., Ishii S., Date T., Sasamoto S., Ishizaki R. Nucleic Acids Res. 16:11075-11090(1988).
- [4] Grotewold E., Athma P., Peterson T. Proc. Natl. Acad. Sci. U.S.A. 88:4587-4591(1991).
- [5] Oppenheimer D.G., Herman P.L., Sivakumaran S., Esch J., Marks M.D. Cell 67:483-493(1991).
- [6] Marocco A., Wissenbach M., Becker D., Paz-Ares J., Saedler H., Salamini F., Rohde W. Mol. Gen. Genet. 216:183-187(1989).
- [7] Tice-Baldwin K., Fink G.R., Arndt K.T. Science 246:931-935(1989).
- [8] Ju Q., Morrow B.E., Warner J.R. Mol. Cell. Biol. 10:5226-5234(1990).
- [9] Klempnauer K.-H., Sippel A.E. EMBO J. 6:2719-2725(1987).

362. NAD-dependent glycerol-3-phosphate dehydrogenase signature

NAD-dependent glycerol-3-phosphate dehydrogenase (EC 1.1.1.8) (GPD) catalyzes the reversible reduction of dihydroxyacetone phosphate to glycerol-3- phosphate. It is a eukaryotic cytosolic homodimeric protein of about 40 Kd. As a signature pattern a glycine-rich region that is probably [1] involved in NAD-binding was selected.

Consensus pattern: G-[AT]-[LIVM]-K-[DN]-[LIVM](2)-A-x-[GA]-x-G-[LIVMF]-x- [DE]-G-[LIVM]-x-[LIVMFYW]-G-x-N-

- [1] Otto J., Argos P., Rossmann M.G. Eur. J. Biochem. 109:325-330(1980).

363. Nucleosome assembly protein (NAP)

It is thought that NAPs may be involved in regulating gene expression as a result of histone accessibility [1].

[1] Rodriguez P, Munroe D, Prawitt D, Chu LL, Bric E, Kim J, Reid LH, Davies C, Nakagama H, Loebbert R, Winterpacht A, Petruzzi MJ, Higgins MJ, Nowak N, Evans G, Shows T, Weissman BE, Zabel B, Housman DE, Pelletier J, Genomics 1997;44:253-265. [2] Schnieders F, Dork T, Arnemann J, Vogel T, Werner M, Schmidtke J; Hum Mol Genet 1996;5:1801-1807.

364. NB-ARC domain

van der Biezen EA, Jones JD, Curr Biol 1998;8:226-227.

365. Nucleoside diphosphate kinases active site

Nucleoside diphosphate kinases (EC 2.7.4.6) (NDK) [1] are enzymes required for the synthesis of nucleoside triphosphates (NTP) other than ATP. They provide NTPs for nucleic acid synthesis, CTP for lipid synthesis, UTP for polysaccharide synthesis and GTP for protein elongation, signal transduction and microtubule polymerization. In eukaryotes, there seems to be a small family of NDK isozymes each of which acts in a different subcellular compartment and/or has a distinct biological function. Eukaryotic NDK isozymes are hexamers of two highly related chains (A and B) [2]. By random association (A₆, A₅B...AB₅, B₆), these two kinds of chain form isoenzymes differing in their isoelectric point. NDK are proteins of 17 Kd that act via a ping-pong mechanism in which a histidine residue is phosphorylated, by transfer of the terminal phosphate group from ATP. In the presence of magnesium, the phosphoenzyme can transfer its phosphate group to any NDP, to produce an NTP. NDK isozymes have been sequenced from prokaryotic and eukaryotic sources. It has also been shown [3] that the *Drosophila* awd (abnormal wing discs) protein, is a microtubule-associated NDK. Mammalian NDK is also known as metastasis inhibition factor nm23. The sequence of NDK has been highly conserved through evolution. There is a single histidine residue conserved in all known NDK isozymes, which is involved in the catalytic mechanism [2]. Our signature pattern contains this residue.

Consensus pattern: N-x(2)-H-[GA]-S-D-[SA]-[LIVMPKNE] [H is the putative active site residue]-

- [1] Parks R., Agarwal R. (In) The Enzymes (3rd edition) 8:307-334(1973).
 [2] Gilles A.-M., Presecan E., Vonica A., Lascu I. J. Biol. Chem. 266:8784-8789(1991).
 [3] Biggs J., Hersperger E., Steeg P.S., Liotta L.A., Shearn A. Cell 63:933-940(1990).

5

366. Nitrite and sulfite reductases iron-sulfur/siroheme-binding site (NIR_SIR)

Nitrite reductases (NiR) [1] catalyze the reduction of nitrite into ammonium, the second step in the assimilation of nitrate. There are two types of NiR: the higher plant chloroplastic form of NiR (EC 1.7.7.1) is a monomeric protein that uses reduced ferredoxin as the electron donor; while fungal and bacterial NiR (EC 1.6.6.4) are homodimeric proteins that uses NAD(P)H as the electron donor. Both forms of NiR contain a siroheme-Fe and iron-sulfur centers. Sulfite reductase (NADPH) (EC 1.8.1.2) (SIR) [2] is the bacterial enzyme that catalyzes the reduction of sulfite to sulfide. SIR is an oligomeric enzyme with a subunit composition of alpha(8)-beta(4), the alpha component is a flavoprotein (SIR-FP), while the beta component is a siroheme, iron-sulfurprotein (SIR-HP). Sulfite reductase (ferredoxin) (EC 1.8.7.1) [3] is a cyanobacterial and plant monomeric enzyme that also catalyzes the reduction of sulfite to sulfide. Anaerobic sulfite reductase (EC 1.8.1.-) (ASR) [4], a bacterial enzyme that catalyzes the NADH-dependent reduction of sulfite to sulfide. ASR is an oligomeric enzyme composed of three different subunits. The C component (geneasrC) seems to be a siroheme, iron-sulfur protein. These enzymes share a region of sequence similarity in their C-terminal half; this region which spans about 80 amino acids includes four conserved cysteine residues. Two of the Cys are grouped together at the beginning of the domain, and the two others are grouped in the middle of the domain. The cysteines are involved in the binding of the iron-sulfur center; the last one also binds the siroheme group [2]. A signature pattern from the region around the second cluster of cysteines was derived.

Consensus pattern: [STV]-G-C-x(3)-C-x(6)-[DE]-[LIVMF]-[GAT]-[LIVMF] [The two C's are iron-sulfur ligands]-

- 30 [1] Campbell W.H., Kinghorn J.R. Trends Biochem. Sci. 15:315-319(1990).
 [2] Crane B.R., Siegel L.M., Getzoff E.D. Science 270:59-67(1995).
 [3] Gisselmann G., Klausmeier P., Schwenn J.D. Biochim. Biophys. Acta 1144:102-106(1993).

[4] Huang C.J., Barrett E.L. J. Bacteriol. 173:1544-1553(1991).

367. (NMT) Myristoyl-CoA:protein N-myristoyltransferase signatures. Myristoyl-CoA:
protein N-myristoyltransferase (EC 2.3.1.97) (Nmt) [1] is the enzyme responsible for
transferring a myristate group on the N-terminal glycine of a number of cellular eukaryotic
and viral proteins. Nmt is a monomeric protein of about 50 to 60 Kd whose sequence appears
to be well conserved. Two highly conserved regions have been developed as signature
patterns. The first one is located in the central section, the second in the C-terminal part.

Consensus pattern: E-I-N-F-L-C-x-H-K-

Consensus pattern: K-F-G-x-G-D-G-

[1] Rudnick D.A., McWherter C.A., Gokel G.W., Gordon J.I. Adv. Enzymol. 67:375-
430(1993).

368. ADP-glucose pyrophosphorylase signatures (NTP_transferase)

ADP-glucose pyrophosphorylase (glucose-1-phosphate adenyltransferase) [1,2](EC
2.7.7.27) catalyzes a very important step in the biosynthesis of alpha 1,4-glucans (glycogen
or starch) in bacteria and plants: synthesis of the activated glucosyl donor, ADP-glucose,
from glucose-1-phosphate and ATP. ADP-glucose pyrophosphorylase is a tetrameric
allosterically regulated enzyme. It is a homotetramer in bacteria while in plant chloroplasts
and amyloplasts, it is a heterotetramer of two different, yet evolutionary related, subunits.

There are a number of conserved regions in the sequence of bacterial and plant ADP-glucose
pyrophosphorylase subunits. Three of these regions were selected as signature patterns. The
first two are N-terminal and have been proposed to be part of the allosteric and/or substrate-
binding sites in the Escherichia coli enzyme (gene glgC). The third pattern corresponds to a
conserved region in the central part of the enzymes.

Consensus pattern: [AG]-G-G-x-G-[STK]-x-L-x(2)-L-[TA]-x(3)-A-x-P-A-[LV] -

Consensus pattern: W-[FY]-x-G-[ST]-A-[DNSH]-[AS]-[LIVMFYW]-

Consensus pattern: [APV]-[GS]-M-G-[LIVMN]-Y-[IVC]-[LIVMFY]-x(2)-[DENPHK] -

[1] Nakata P.A., Greene T.W., Anderson J.M., Smith-White B.J., Okita T.W., Preiss J. Plant Mol. Biol. 17:1089-1093(1991).

[2] Preiss J., Ball K., Hutney J., Smith-White B.J., Li. L., Okitsa T.W. Pure Appl. Chem. 63:535-544(1991).

369. Sodium/hydrogen exchanger family

Na/H antiporters are key transporters in maintaining the pH of actively metabolizing cells. The molecular mechanisms of antiport are unclear.

These antiporters contain 10-12 transmembrane regions (M) at the amino-terminus and a large cytoplasmic region at the carboxyl terminus. The transmembrane regions M3-M12 share identity with other members of the family. The M6 and M7 regions are highly conserved. Thus, this is thought to be the region that is involved in the transport of sodium and hydrogen ions. The cytoplasmic region has little similarity throughout the family.

[1] Dibrov P, Fliegel L; FEBS Lett 1998;424:1-5. [2] Orlowski J, Grinstein S; J Biol Chem 1997;272:22373-22376.[3] Numata M, Petrecca K, Lake N, Orlowski J; J Biol Chem 1998;273:6951-6959.

370. Sodium:sulfate symporter family signature (Na_sulph_symp)

Integral membrane proteins that mediate the intake of a wide variety of molecules with the concomitant uptake of sodium ions (sodium symporters) can be grouped, on the basis of sequence and functional similarities into a number of distinct families. One of these families currently consists of the following proteins: - Mammalian sodium/sulfate cotransporter [1]. - Mammalian renal sodium/dicarboxylate cotransporter [2], which transports succinate and citrate. - Mammalian intestinal sodium/dicarboxylate cotransporter. - Chlamydomonas reinhardtii putative sulfur deprivation response regulator SAC1 [3]. - Caenorhabditis elegans hypothetical proteins B0285.6, F31F6.6, K08E5.2 and R107.1. - Escherichia coli hypothetical protein yfbS. - Haemophilus influenzae hypothetical protein HI0608. - Synechocystis strain

359

PCC 6803 hypothetical protein sll0640. - Methanococcus jannaschii hypothetical protein MJ0672. These transporters are proteins of from 430 to 620 amino acids which are highly hydrophobic and which probably contain about 12 transmembrane regions. As a signature pattern, a conserved region was selected which is located in or near the penultimate transmembrane region.

Consensus pattern: [STACP]-S-x(2)-F-x(2)-P-[LIVM]-[GSA]-x(3)-N-x-[LIVM]-V-

[1] Markovich D., Forgo J., Stange G., Biber J., Murer H. Proc. Natl. Acad. Sci. U.S.A. 90:8073-8077(1993).

[2] Pajor A.M. Am. J. Physiol. 270:642-648(1996).

[3] Davies J.P., Yildiz F.H., Grossman A. EMBO J. 15:2150-2159(1996).

371. NifU-like domain

This is an alignment of the carboxy-terminal domain. This is the only common region between the NifU protein from nitrogen-fixing bacteria and rhodobacterial species. The biochemical function of NifU is unknown [1].

Ouzounis C, Bork P, Sander C, Trends Biochem Sci 1994;19:199-200.

372. Nitrilases / cyanide hydratase signatures

Nitrilases (EC 3.5.5.1) are enzymes that convert nitriles into their corresponding acids and ammonia. They are widespread in microbes as well as in plants where they convert indole-3-acetonitrile to the hormone indole-3-acetic acid. A conserved cysteine has been shown [1,2] to be essential for enzyme activity; it seems to be involved in a nucleophilic attack on the nitrile carbon atom. Cyanide hydratase (EC 4.2.1.66) converts HCN to formamide. In phytopathogenic fungi, it is used to avoid the toxic effect of cyanide released by wounded plants [3]. The sequence of cyanide hydrolase is evolutionary related to that of nitrilases.

Yeast hypothetical proteins YIL164c and YIL165c also belong to this family. As signature patterns for these enzymes, two conserved regions were selected. The first is located in the N-terminal section while the second, which contains the active site cysteine, is located in the central section.

Consensus pattern: G-x(2)-[LIVMFY](2)-x-[IF]-x-E-x(2)-[LIVM]-x-G-Y-P-

Consensus pattern: G-[GAQ]-x(2)-C-[WA]-E-[NH]-x(2)-[PST]-[LIVMFYS]-x-[KR] [C is the active site residue]-

5

[1] Kobayashi M., Izui H., Nagasawa T., Yamada H. Proc. Natl. Acad. Sci. U.S.A. 90:247-251(1993).

[2] Kobayashi M., Komeda H., Yanaka N., Nagasawa T., Yamada H. J. Biol. Chem. 267:20746-20751(1992).

10 [3] Wang P., Vanetten H.D. Biochem. Biophys. Res. Commun. 187:1048-1054(1992).

373. NusB family

The NusB protein is involved in the regulation of rRNA biosynthesis by transcriptional antitermination.

15

Huenges M, Rolz C, Gschwind R, Peteranderl R, Berglechner F, Richter G, Bacher A, Kessler H, Gemmecker G, EMBO J 1998;17:4092-4100.

374. (Neur Chan) Neurotransmitter-gated ion-channels signature

Neurotransmitter-gated ion-channels [1,2,3,4] provide the molecular basis for rapid signal transmission at chemical synapses. They are post-synaptic oligomeric transmembrane complexes that transiently form an ionic channel upon the binding of a specific neurotransmitter. Presently, the sequence of subunits from five types of neurotransmitter-gated receptors are known: - The nicotinic acetylcholine receptor (AChR), an excitatory cation channel. In the motor endplates of vertebrates, it is composed of four different subunits (alpha, beta, gamma and delta or epsilon) with a molar stoichiometry of 2:1:1:1. In neurones, the AChR receptor is composed of two different types of subunits: alpha and non-alpha (also called beta). Nicotinic AChRs are also found in invertebrates. - The glycine receptor, an inhibitory chloride ion channel. The glycine receptor is a pentamer composed of two different subunits (alpha and beta). - The gamma-aminobutyric-acid (GABA) receptor, which is also an inhibitory chloride ion channel. The quaternary structure of the GABA receptor is complex; at least four classes of subunits are known to exist (alpha, beta, gamma, and delta)

25

30

and there are many variants in each class (for example: six variants of the alpha class have already been sequenced). - The serotonin 5HT3 receptor. Serotonin is a biogenic hormone that functions as a neurotransmitter, a hormone and a mitogen. There are seven major groups of serotonin receptors; six of these groups (5HT1, 5HT2, and 5HT4 to 5HT7) transduce extracellular signal by activating G proteins, while 5HT3 is a ligand-gated cation-specific ion channel which, when activated causes fast, depolarizing responses in neurons. - The glutamate receptor, an excitatory cation channel. Glutamate is the main excitatory neurotransmitter in the brain. At least three different types of glutamate receptors have been described and are named according to their selective agonists (kainate, N-methyl-D-aspartate (NMDA) and quisqualate). All known sequences of subunits from neurotransmitter-gated ion-channels are structurally related. They are composed of a large extracellular glycosylated N-terminal ligand-binding domain, followed by three hydrophobic transmembrane regions which form the ionic channel, followed by an intracellular region of variable length. A fourth hydrophobic region is found at the C-terminal of the sequence. The sequence of subunits from the AchR, GABA, 5HT3, and Gly receptors are clearly evolutionary related and share many regions of sequence similarities. These sequence similarities are either absent or very weak in the Glu receptors. In the N-terminal extracellular domain of AchR/GABA/5HT3/Gly receptors, there are two conserved cysteine residues, which, in AchR, have been shown to form a disulfide bond essential to the tertiary structure of the receptor. A number of amino acids between the two disulfide-bonded cysteines are also conserved. Therefore this region was used as a signature pattern for this subclass of proteins.

Consensus pattern: C-x-[LIVMFQ]-x-[LIVMF]-x(2)-[FY]-P-x-D-x(3)-C [The two C's are linked by a disulfide bond]-

- [1] Stroud R.M., McCarthy M.P., Shuster M. Biochemistry 29:11009-11023(1990).
- [2] Betz H. Neuron 5:383-392(1990).
- [3] Dingledine R., Myers S.J., Nicholas R.A. FASEB J. 4:2632-2645(1990).
- [4] Barnard E.A. Trends Biochem. Sci. 17:368-374(1992).

375. Orotidine 5'-phosphate decarboxylase active site

Orotidine 5'-phosphate decarboxylase (EC 4.1.1.23) (OMPdecase) [1,2] catalyzes the last step in the de novo biosynthesis of pyrimidines, the decarboxylation of OMP into UMP. In higher eukaryotes OMPdecase is part, with orotatephosphoribosyltransferase, of a bifunctional enzyme, while the prokaryotic and fungal OMPdecases are monofunctional protein. Some parts of the sequence of OMPdecase are well conserved across species. The best conserved region is located in the N-terminal half of OMPdecases and is centered around a lysine residue which is essential for the catalytic function of the enzyme. This region has been developed as a signature pattern.

Consensus pattern: [LIVMFTA]-[LIVMF]-x-D-x-K-x(2)-D-I-[GP]-x-T-[LIVMTA] [K is the active site residue]-

[1] Jacquet M., Guilbaud R., Garreau H. Mol. Gen. Genet. 211:441-445(1988).

[2] Kimsey H.H., Kaiser D. J. Biol. Chem. 267:819-824(1992).

376. ATP synthase delta (OSCP) subunit signature

ATP synthase (proton-translocating ATPase) (EC 3.6.1.34) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria, and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), which acts as a proton channel, and a catalytic core, termed coupling factor CF(1).

One of the subunits of the ATPase complex, known as subunit delta in bacteria and chloroplasts or the Oligomycin Sensitivity Conferral Protein (OSCP) in mitochondria, seems to be part of the stalk that links CF(0) to CF(1). It either transmits conformational changes from CF(0) into CF(1) or is involved in proton conduction [3].

The different delta/OSCP subunits are proteins of approximately 200 amino-acid residues - once the transit peptide has been removed in the chloroplast and mitochondrial forms - which show only moderate sequence homology.

The signature pattern used to detect ATPase delta/OSCP subunits is based on a conserved region in the C-terminal section of these proteins.

Consensus pattern: [LIVM]-x-[LIVMFYT]-x(3)-[LIVMT]-[DENQK]-x(2)-[LIVM]-x-[GSA]-G-[LIVMFYGA]-x-[LIVM]-[KRHENQ]-x-[GSEN]

[1] Futai M., Noumi T., Maeda M. Annu. Rev. Biochem. 58:111-136(1989).

5 [2] Senior A.E. Physiol. Rev. 68:177-231(1988).

[3] Engelbrecht S., Junge W. Biochim. Biophys. Acta 1015:379-390(1990).

377. Aspartate and ornithine carbamoyltransferases signature

10 Aspartate carbamoyltransferase (EC 2.1.3.2) (ATCase) catalyzes the conversion of aspartate and carbamoyl phosphate to carbamoylaspartate, the second step in the de novo biosynthesis of pyrimidine nucleotides [1]. In prokaryotes ATCase consists of two subunits: a catalytic chain (gene pyrB) and a regulatory chain (gene pyrI), while in eukaryotes it is a domain in a multi-
15 functional enzyme (called URA2 in yeast, rudimentary in Drosophila, and CAD in mammals [2]) that also catalyzes other steps of the biosynthesis of pyrimidines.

Ornithine carbamoyltransferase (EC 2.1.3.3) (OTCase) catalyzes the conversion of ornithine and carbamoyl phosphate to citrulline. In mammals this enzyme
20 participates in the urea cycle [3] and is located in the mitochondrial matrix. In prokaryotes and eukaryotic microorganisms it is involved in the biosynthesis of arginine. In some bacterial species it is also involved in the degradation of arginine [4] (the arginine deaminase pathway).

It has been shown [5] that these two enzymes are evolutionary related. The
25 predicted secondary structure of both enzymes are similar and there are some regions of sequence similarities. One of these regions includes three residues which have been shown, by crystallographic studies [6], to be implicated in binding the phosphoryl group of carbamoyl phosphate.

This region was selected as a signature for these enzymes.

30

Consensus pattern: F-x-[EK]-x-S-[GT]-R-T[S, R, and the 2nd T bind carbamoyl phosphate]

-Note: the residue in position 3 of the pattern allows to distinguish between an ATCase (Glu) and an OTCase (Lys).

- [1] Lerner C.G., Switzer R.L. J. Biol. Chem. 261:11156-11165(1986).
- [2] Davidson J.N., Chen K.C., Jamison R.S., Musmanno L.A., Kern C.B. BioEssays 15:157-164(1993).
- 5 [3] Takiguchi M., Matsubasa T., Amaya Y., Mori M. BioEssays 10:163-166(1989).
- [4] Baur H., Stalon V., Falmagne P., Luethi E., Haas D. Eur. J. Biochem. 166:111-117(1987).
- [5] Houghton J.E., Bencini D.A., O'Donovan G.A., Wild J.R. Proc. Natl. Acad. Sci. U.S.A. 81:4864-4868(1981).
- 10 [6] Ke H.-M., Honzatko R.B., Lipscomb W.N. Proc. Natl. Acad. Sci. U.S.A. 81:4037-4040(1984).

378. Oleosins signature

15 Oleosins [1] are the proteinaceous components of plants' lipid storage bodies called oil bodies. Oil bodies are small droplets (0.2 to 1.5 μ m in diameter) containing mostly triacylglycerol that are surrounded by a phospholipid/oleosin annulus. Oleosins may have a structural role in stabilizing the lipid body during dessication of the seed, by preventing coalescence of the oil.

20 They may also provide recognition signals for specific lipase anchorage in lipolysis during seedling growth. Oleosins are found in the monolayer lipid/water interface of oil bodies and probably interact with both the lipid and phospholipid moieties.

Oleosins are proteins of 16 Kd to 24 Kd and are composed of three domains: an

25 N-terminal hydrophilic region of variable length (from 30 to 60 residues); a central hydrophobic domain of about 70 residues and a C-terminal amphipathic region of variable length (from 60 to 100 residues). The central hydrophobic domain is proposed to be made up of beta-strand structure and to interact with the lipids [2]. It is the only domain whose sequence is conserved and therefore

30 a section from that domain was selected as a signature pattern.

Consensus pattern: [AG]-[ST]-x(2)-[AG]-x(2)-[LIVM]-[SAD]-T-P-[LIVMF](4)-F-S-P-[LIVM](3)-P-A

- [1] Murphy D.J., Keen J.N., O'Sullivan J.N., Au D.M.Y., Edwards E.-W., Jackson P.J., Cummins I., Gibbons T., Shaw C.H., Ryan A.J. *Biochim. Biophys. Acta* 1088:86-94(1991).
 [2] Tzen J.T.C., Lie G.C., Huang A.H.C. *J. Biol. Chem.* 267:15626-15634(1992).

5

379. (Orbi VP5) Orbivirus outer capsid protein VP5

This paper shows the location of the different capsid proteins
 and their relation to each other.

10

- [1] Schoehn G, Moss SR, Nuttall PA, Hewat EA; *Virology* 1997;235:191-200.

380. Orn/DAP/Arg decarboxylases family 2 signatures

Pyridoxal-dependent decarboxylases acting on ornithine, lysine, arginine and related substrates can be classified into two different families on the basis of sequence similarities [1,2,3]. The second family consists of:

- Eukaryotic ornithine decarboxylase (EC 4.1.1.17) (ODC). ODC catalyzes the transformation of ornithine into putrescine.
- Prokaryotic diaminopimelic acid decarboxylase (EC 4.1.1.20) (DAPDC). DAPDC catalyzes the conversion of diaminopimelic acid into lysine; the last step in the biosynthesis of lysine.
- *Pseudomonas syringae* pv. *tabaci* protein tabA. tabA is probably involved in the biosynthesis of tabtoxin and is highly similar to DAPDC.
- Bacterial and plant biosynthetic arginine decarboxylase (EC 4.1.1.19) (ADC). ADC catalyzes the transformation of arginine into agmatine, the first step in the biosynthesis of putrescine from arginine.

The above proteins, while most probably evolutionary related, do not share extensive regions of sequence similarities. Two of the conserved regions were selected as signature patterns. The first pattern contains a conserved lysine residue which is known, in mouse ODC [4], to be the site of attachment of the pyridoxal-phosphate group. The second pattern contains a stretch of three

30

consecutive glycine residues and has been proposed to be part of a substrate-binding region [5].

These enzymes are collectively known as group IV decarboxylases [3].

- 5 Consensus pattern: [FY]-[PA]-x-K-[SACV]-[NHCLFW]-x(4)-[LIVMF]-[LIVMTA]-x(2)-
[LIVMA]-x(3)-[GTE] [K is the pyridoxal-P attachment site]
Consensus pattern: [GS]-x(2,6)-[LIVMSCP]-x(2)-[LIVMF]-[DNS]-[LIVMCA]-G-G-G-
[LIVMFY]-[GSTPCEQ]

- 10 [1] Bairoch A. Unpublished observations (1993).
[2] Martin C., Cami B., Yeh P., Stragier P., Parsot C., Patte J.-C. Mol. Biol. Evol. 5:549-
559(1988).
[3] Sandmeier E., Hale T.I., Christen P. Eur. J. Biochem. 221:997-1002(1994).
[4] Poulin R., Lu L., Ackermann B., Bey P., Pegg A.E. J. Biol. Chem. 267:150-158(1992).
15 [5] Moore R.C., Boyle S.M. J. Bacteriol. 172:4631-4640(1990).

381. Osteopontin signature

Osteopontin is an acidic phosphorylated glycoprotein of about 40 Kd which is
20 abundant in the mineral matrix of bones and which binds tightly to
hydroxyapatite [1,2,3]. It is suggested that osteopontin might function as a
cell attachment factor and could play a key role in the adhesion of
osteoclasts to the mineral matrix of bone.

Osteopontin-K is a kidney protein which is highly similar to osteopontin and
25 probably also involved in cell-adhesion.

As a signature pattern a highly conserved region located at the
N-terminal extremity of the mature protein was selected.

Consensus pattern: [KQ]-x-[TA]-x(2)-[GA]-S-S-E-E-K

- 30 [1] Butler W.T. Connect. Tissue Res. 23:123-36(1989).
[2] Gorski J.P. Calcif. Tissue Int. 50:391-396(1992).
[3] Denhardt D.T., Guo X. FASEB J. 7:1475-1482(1993).

382. Oxysterol-binding protein family signature

A number of eukaryotic proteins that seem to be involved with sterol synthesis and/or its regulation have been found [1] to be evolutionary related:

- Mammalian oxysterol-binding protein (OSBP). A protein of about 800 amino-acid residues that binds a variety of oxysterols: oxygenated derivatives of cholesterol. OSBP seems to play a complex role in the regulation of sterol metabolism.
- Yeast proteins HES1 and KES1; highly related proteins of 434 residues that seem to play a role in ergosterol synthesis.
- Yeast OSH1, a protein of 859 residues that also plays a role in ergosterol synthesis.
- Yeast hypothetical protein YHR001w (437 residues).
- Yeast hypothetical protein YHR073w (996 residues).
- Yeast hypothetical protein YKR003w (448 residues).

All these proteins contain a moderately conserved domain of about 250 residues located in the C-terminal half of OBSP, OSH1 and YHR073w and in the central section of the other proteins. As a signature pattern, the best conserved part was selected of this domain, a region that contains a conserved pentapeptide.

Consensus pattern: E-[KQ]-x-S-H-[HR]-P-P-x-[STACF]-A

[1] Jiang B., Brown J.L., Sheraton J., Fortin N., Bussey H. Yeast 10:341-353(1994).

383. FMN oxidoreductase

384. Oxidoreductase FAD/NAD-binding domain

Number of members: 250

[1]

Medline: 92084635

The sequence of squash NADH:nitrate reductase and its relationship to the sequences of other flavoprotein oxidoreductases. A family of flavoprotein pyridine nucleotide cytochrome reductases.

- 5 Hyde GE, Crawford NM, Campbell W;
J Biol Chem 1991;266:23542-23547.
[2]Medline: 95111952

Crystal structure of the FAD-containing fragment of corn nitrate reductase at 2.5 Å resolution: relationship to other flavoprotein reductases.

- 10 Lu G, Campbell WH, Schneider G, Lindqvist Y;
Structure 1994;2:809-821.

385. (oxidored molyb) Eukaryotic molybdopterin oxidoreductases signature
A number of different eukaryotic oxidoreductases that require and bind a molybdopterin cofactor have been shown [1] to share a few regions of sequence similarity. These enzymes are:

- 15 - Xanthine dehydrogenase (EC 1.1.1.204), which catalyzes the oxidation of xanthine to uric acid with the concomitant reduction of NAD. Structurally, this enzyme of about 1300 amino acids consists of at least three distinct domains: an N-terminal 2Fe-2S ferredoxin-like iron-sulfur binding domain (see <PDOC00175>), a central FAD/NAD-binding domain and a C-terminal Molybdopterin domain.
- 20 - Aldehyde oxidase (EC 1.2.3.1), which catalyzes the oxidation aldehydes into acids. Aldehyde oxidase is highly similar to xanthine dehydrogenase in its sequence and domain structure.
- 25 - Nitrate reductase (EC 1.6.6.1), which catalyzes the reduction of nitrate to nitrite. Structurally, this enzyme of about 900 amino acids consists of an N-terminal Molybdopterin domain, a central cytochrome b5-type heme-binding domain (see <PDOC00170>) and a C-terminal FAD/NAD-binding cytochrome reductase domain.
- 30 - Sulfite oxidase (EC 1.8.3.1), which catalyzes the oxidation of sulfite to

369

sulfate. Structurally, this enzyme of about 460 amino acids consists of an N-terminal cytochrome b5-binding domain followed by a Mo-pterin domain. There are a few conserved regions in the sequence of the molybdopterin-binding domain of these enzymes. The pattern used to detect these proteins is based on one of them. It contains a cysteine residue which could be involved in binding the molybdopterin cofactor.

Consensus pattern: [GA]-x(3)-[KRNQHT]-x(11,14)-[LIVMFYWS]-x(8)-[LIVMF]-x-C-x(2)-[DEN]-R-x(2)-[DE]

[1] Wootton J.C., Nicolson R.E., Cock J.M., Walters D.E., Burke J.F., Doyle W.A., Bray R.C. Biochim. Biophys. Acta 1057:157-185(1991).

386. (Oxidored q1) NADH-Ubiquinone/plastoquinone (complex I), various chains

This family is part of complex I which catalyses the transfer of two electrons from NADH to ubiquinone in a reaction that is associated with proton translocation across the membrane. Number of members: 1824

[1]

Medline: 93110040

The NADH:ubiquinone oxidoreductase (complex I) of respiratory chains. Walker JE; Q Rev Biophys 1992;25:253-324.

387. (oxidored q3) NADH-ubiquinone/plastoquinone oxidoreductase chain 6. 179 members.

388. (oxidored q5) NADH-ubiquinone oxidoreductase chain 4, amino terminus

[1] Walker JE ; Q Rev Biophys 1992;25:253-324.

370

389. (oxidored q6) Respiratory-chain NADH dehydrogenase 20 Kd subunit signature
Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex
I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex
located in the inner mitochondrial membrane which also seems to exist in
the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase).
Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex
there is one with a molecular weight of 20 Kd (in mammals) [3], which is a
component of the iron-sulfur (IP) fragment of the enzyme. It seems to bind a
4Fe-4S iron-sulfur cluster. The 20 Kd subunit has been found to be:

- Nuclear encoded, as a precursor form with a transit peptide in mammals, and
in *Neurospora crassa*. - Mitochondrial encoded in *Paramecium* (gene *psbG*).
- Chloroplast encoded in various higher plants (gene *ndhK* or *psbG*).

The 20 Kd subunit is highly similar to [4]:

- *Synechocystis* strain PCC 6803 proteins *psbG1* and *psbG2*.
- Subunit B of *Escherichia coli* NADH-ubiquinone oxidoreductase (gene *nuoB*).
- Subunit NQO6 of *Paracoccus denitrificans* NADH-ubiquinone oxidoreductase.
- Subunit 7 of *Escherichia coli* formate hydrogenlyase (gene *hycG*).
- Subunit I of *Escherichia coli* hydrogenase-4 (gene *hyfI*).

As as signature pattern a highly conserved region was selected, located in the
central section of this subunit and which contains a conserved cysteine that
is probably involved in the binding of the 4Fe-4S center.

Consensus pattern: [GN]-x-D-[KRST]-[LIVMF](2)-P-[IV]-D-[LIVMFYW](2)-x-P-x-C-P-
[PT] [The C is a putative 4Fe-4S ligand]

- [1] Ragan C.I. Curr. Top. Bioenerg. 15:1-36(1987).
- [2] Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).
- [3] Arizmendi J.M., Runswick M.J., Skehel J.M., Walker J.E. FEBS Lett. 301:237-
242(1992).
- [4] Weidner U., Geier S., Ptock A., Friedrich T., Leif H., Weiss H. J. Mol. Biol. 233:109-
122(1993).

390. p53 tumor antigen signature

The p53 tumor antigen [1 to 5, E1,E2] is a protein found in increased amounts in a wide variety of transformed cells. It is also detectable in many proliferating nontransformed cells, but it is undetectable or present at low levels in resting cells. It is frequently mutated or inactivated in many types of cancer. p53 seems to act as a tumor suppressor in some, but probably not all, tumor types. p53 is probably involved in cell cycle regulation, and may be a trans-activator that acts to negatively regulate cellular division by controlling a set of genes required for this process.

p53 is a phosphoprotein of about 390 amino acids which can be subdivided into four domains: a highly charged acidic region of about 75 to 80 residues, a hydrophobic proline-rich domain (position 80 to 150), a central region (from 150 to about 300), and a highly basic C-terminal region. The sequence of p53 is well conserved in vertebrate species; attempts to identify p53 in other eukaryotic phylum has so far been unsuccessful.

As a signature pattern for p53 a perfectly conserved stretch of 13 residues located in the central region of the protein was selected. This region, known as domain IV in [3], is involved (along with an adjacent region) in the binding of the large T antigen of SV40. In man this region is the focus of a variety of point mutations in cancerous tumors.

Consensus pattern: M-C-N-S-S-C-M-G-G-M-N-R-R

[1] Levine A.J., Momand J., Finlay C.A. Nature 351:453-456(1991).

[2] Levine A.J., Momand J. Biochim. Biophys. Acta 1032:119-136(1990).

[3] Soussi T., Caron De Fromental C., May P. Oncogene 5:945-952(1990).

[4] Lane D.P., Benchimol S. Genes Dev. 4:1-8(1990).

[5] Ulrich S.J., Anderson C.W., Mercer W.E., Appella E. J. Biol. Chem. 267:15259-15262(1992).

391. (P5CR) Delta 1-pyrroline-5-carboxylate reductase signature

Delta 1-pyrroline-5-carboxylate reductase (P5CR) (EC 1.5.1.2) [1,2] is the enzyme that catalyzes the terminal step in the biosynthesis of proline from glutamate, the NAD(P) dependent oxidation of 1-pyrroline-5-carboxylate into

proline.

The sequences of P5CR from eubacteria (gene proC), archaeobacteria and eukaryotes show only a moderate level of overall similarity. As a signature pattern, the best conserved region located in the C-terminal

5 section of P5CR was selected.

Consensus pattern: [PALF]-x(2,3)-[LIV]-x(3)-[LIVM]-[STAC]-[STV]-x-[GAN]-G-x-T-x(2)-[AG]-[LIV]-x(2)-[LMF]-[DENQK]

10 [1] Delauney A.J., Verma D.P. Mol. Gen. Genet. 221:299-305(1990).

[2] Savioz A., Jeenes D.J., Kocher H.P., Haas D. Gene 86:107-111(1990).

392. Poly-adenylate binding protein, unique domain.

393. (PAL) Phenylalanine and histidine ammonia-lyases active site

Phenylalanine ammonia-lyase (EC 4.3.1.5) (PAL) is a key enzyme of plant and fungi phenylpropanoid metabolism which is involved in the biosynthesis of a wide variety of secondary metabolites such as flavanoids, furanocoumarin phytoalexins and cell wall components. These compounds have many important roles in plants during normal growth and in responses to environmental stress. PAL catalyzes the removal of an ammonia group from phenylalanine to form trans-cinnamate.

25 Histidine ammonia-lyase (EC 4.3.1.3) (histidase) catalyzes the first step in histidine degradation, the removal of an ammonia group from histidine to produce urocanic acid.

The two types of enzymes are functionally and structurally related [1]. They are the only enzymes which are known to have the modified amino acid dehydro-alanine (DHA) in their active site. A serine residue has been shown [2,3,4] to be the precursor of this essential electrophilic moiety. The region around this active site residue is well conserved and can be used as a signature pattern.

Consensus pattern: G-[STG]-[LIVM]-[STG]-[AC]-S-G-[DH]-L-x-P-L-[SA]-x(2)-[SA] [S is the active site residue]

- 5 [1] Taylor R.G., Lambert M.A., Sexsmith E., Sadler S.J., Ray P.N., Mahuran D.J., McInnes R.R. J. Biol. Chem. 265:18192-18199(1990).
- [2] Langer M., Reck G., Reed J., Retey J. Biochemistry 33:6462-6467(1994).
- [3] Schuster B., Retey J. FEBS Lett. 349:252-254(1994).
- [4] Taylor R.G., McInnes R.R. J. Biol. Chem. 269:27473-27477(1994).

10

394. PAS domain

-!- CAUTION. This family does not currently match all known examples of PAS domains.

PAS motifs appear in archaea, eubacteria and eukarya. Probably the most surprising identification of a PAS domain was that in EAG-like K⁺-channels[1,3].

Number of members: 308

[1]

Medline: 97446881

PAS domain S-boxes in archaea, bacteria and sensors for oxygen and redox.

Zhulin IB, Taylor BL, Dixon R;

Trends Biochem Sci 1997;22:331-333.

[2]Medline: 95275818

1.4 A structure of photoactive yellow protein, a cytosolic photoreceptor: unusual fold, active site, and chromophore.

Borgstahl GE, Williams DR, Getzoff ED;

Biochemistry 1995;34:6278-6287.

[3]Medline: 98044337

PAS: a multifunctional domain family comes to light.

Ponting CP, Aravind L;

Curr Biol 1997;7:674-677.

25

30

15

395. (PBP) Phosphatidylethanolamine-binding protein family signature

Mammalian phosphatidylethanolamine-binding protein (also known as basic cytosolic 21 Kd protein) is a 186 residue protein found in a variety of tissues [1]. It binds hydrophobic ligands, such as phosphatidylethanolamine, but also seems [2] to bind nucleotides such as GTP and FMN, it is suggested that it could act in membrane remodeling during growth and maturation. This protein belongs to a family that also includes:

- Drosophila antennal protein A5, a putative odorant-binding protein.
- Onchocerca volvulus antigen Ov-16 and the related proteins D1, D2 and D3.
- Plasmodium falciparum putative phosphatidylethanolamine-binding protein.
- Toxocara canis secreted antigen TES-26. This larval protein has been shown to bind phosphatidylethanolamine.
- Yeast protein DKA1 (also known as NSP1 or TFS1). The function of this protein is not very clear.
- Yeast hypothetical protein YLR179C.
- Caenorhabditis elegans hypothetical protein F40A3.3.

As a signature pattern, the best conserved region was selected which is located in the end of the first third of the sequence of these proteins.

Consensus pattern: [FYL]-x-[LV]-[LIVF]-x-[TIV]-[DC]-P-D-x-P-[SN]-x(10)-H

[1] Seddiqi N., Bollengier F., Alliel P.M., Perin J.P., Bonnet F., Bucquoy S., Jolles P., Schoentgen F. J. Mol. Evol. 39:655-660(1994).

[2] Schoentgen F., Jolles P. FEBS Lett. 369:22-6(1995).

396. PCI domain

This domain has also been called the PINT motif (Proteasome, Int-6, Nip-1 and TRIP-15) [1].

Number of members: 49

[1]

Medline: 98308842

The PCI domain: a common theme in three multiprotein complexes.

Hofmann K, Bucher P;

Trends Biochem Sci 1998;23:204-205.

[2]Medline: 98266368

Homologues of 26S proteasome subunits are regulators of transcription and translation.

Aravind L, Ponting CP;

Protein Sci 1998;7:1250-1254.

397. (PCMT) Protein-L-isoaspartate (D-aspartate) O-methyltransferase signature. Protein-L-isoaspartate (D-aspartate) O-methyltransferase (EC 2.1.1.77) (PCMT)[1] (which is also known as L-isoaspartyl protein carboxyl methyltransferase) is an enzyme that catalyzes the transfer of a methyl group from S-adenosylmethionine to the free carboxyl groups of D-aspartyl or L-isoaspartyl residues in a variety of peptides and proteins. The enzyme does not act on normal L-aspartyl residues L-isoaspartyl and D-aspartyl are the products of the spontaneous de amidation and/or isomerization of normal L-aspartyl and L-asparaginyl residues in proteins. PCMT plays a role in the repair and/or degradation of these damaged proteins; the enzymatic methyl esterification of the abnormal residues can lead to their conversion to normal L-aspartyl residues. PCMT is a well-conserved and widely distributed cytosolic protein of about 24Kd. As a signature pattern, a conserved region in the central part of this enzyme has been developed.

Consensus pattern: [GSA]-D-G-x(2)-G-[FYWV]-x(3)-[AS]-P-[FY]-[DN]-x-I -

[1] Kagan R.M., McFadden H.J., McFadden P.N., O'Connor C., Clarke S. Comp. Biochem. Physiol. 117b:379-385(1997).

398. (PCNA) Proliferating cell nuclear antigen signatures

Proliferating cell nuclear antigen (PCNA) [1,2] is a protein involved in DNA replication by acting as a cofactor for DNA polymerase delta, the

polymerase responsible for leading strand DNA replication.

A similar protein exists in yeast (gene POL30) [3] and is associated with polymerase III, the yeast analog of polymerase delta. In baculoviruses the ETL protein has been shown [4] to be highly related to PCNA and is probably associated with the viral encoded DNA polymerase. An homolog of PCNA is also found in archebacteria.

As signatures for this family of proteins, two conserved regions were selected located in the N-terminal section. The second one has been proposed to bind DNA.

Consensus pattern: [GA]-[LIVMF]-x-[LIVMA]-x-[SAV]-[LIVM]-D-x-[NSAE]-[HKR]-[VI]-x-[LY]-[VGA]-x-[LIVM]-x-[LIVM]-x(4)-F
-Consensus pattern: [RKA]-C-[DE]-[RH]-x(3)-[LIVMF]-x(3)-[LIVM]-x-[SGAN]-[LIVMF]-x-K-[LIVMF](2)

- [1] Bravo R., Frank R., Blundell P.A., McDonald-Bravo H. Nature 326:515-517(1987).
- [2] Suzuka I., Hata S., Matsuoka M., Kosugi S., Hashimoto J. Eur. J. Biochem. 195:571-575(1991).
- [3] Bauer G.A., Burgess P.M.J. Nucleic Acids Res. 18:261-265(1990).
- [4] O'Reilly D.R., Crawford A.M., Miller L.K. Nature 337:606-606(1989).

399. (PDT) Prephenate dehydratase signatures

Prephenate dehydratase (EC 4.2.1.51) (PDT) catalyzes the decarboxylation of prephenate into phenylpyruvate. In microorganisms PDT is involved in the terminal pathway of the biosynthesis of phenylalanine. In some bacteria such as Escherichia coli PDT is part of a bifunctional enzyme (P-protein) that also catalyzes the transformation of chorismate into prephenate (chorismate mutase) while in other bacteria it is a monofunctional enzyme. The sequence of monofunctional PDT align well with the C-terminal part of that of P-proteins [1].

As signature patterns for PDT two conserved regions were selected. The first region contains a conserved threonine which has been said to be essential for the activity of the enzyme in E. coli. The second region includes a conserved

glutamate. Both regions are in the C-terminal part of PDT.

Consensus pattern: [FY]-x-[LIVM]-x(2)-[LIVM]-x(5)-[DN]-x(5)-T-R-F-[LIVMW]-x-[LIVM]

5

[1] Fischer R.S., Zhao G., Jensen R.A. J. Gen. Microbiol. 137:1293-1301(1991).

400. PDZ domain (Also known as DHR or GLGF).

10

PDZ domains are found in diverse signaling proteins.

[1] Ponting CP, Phillips C, Davies KE, Blake DJ

Bioessays 1997;19:469-479. [2] Doyle DA, Lee A, Lewis J, Kim E, Sheng M, MacKinnon R; Cell. 1996;85:1067-1076. [3] Ponting CP; Protein Sci 1997;6:464-468.

15

401. (PPDK_N_term) PEP-utilizing enzymes signatures

A number of enzymes that catalyze the transfer of a phosphoryl group from phosphoenolpyruvate (PEP) via a phospho-histidine intermediate have been shown to be structurally related [1,2,3,4]. These enzymes are:

20

- Pyruvate,orthophosphate dikinase (EC 2.7.9.1) (PPDK). PPDK catalyzes the reversible phosphorylation of pyruvate and phosphate by ATP to PEP and diphosphate. In plants PPDK function in the direction of the formation of PEP, which is the primary acceptor of carbon dioxide in C4 and crassulacean acid metabolism plants. In some bacteria, such as Bacteroides symbiosus, PPDK functions in the direction of ATP synthesis.

25

- Phosphoenolpyruvate synthase (EC 2.7.9.2) (pyruvate,water dikinase). This enzyme catalyzes the reversible phosphorylation of pyruvate by ATP to form PEP, AMP and phosphate, an essential step in gluconeogenesis when pyruvate and lactate are used as a carbon source.

30

- Phosphoenolpyruvate-protein phosphotransferase (EC 2.7.3.9). This is the first enzyme of the phosphoenolpyruvate-dependent sugar phosphotransferase system (PTS), a major carbohydrate transport system in bacteria. The PTS catalyzes the phosphorylation of incoming sugar substrates concomitant

with their translocation across the cell membrane. The general mechanism of the PTS is the following: a phosphoryl group from PEP is transferred to enzyme-I (EI) of PTS which in turn transfers it to a phosphoryl carrier protein (HPr). Phospho-HPr then transfers the phosphoryl group to a sugar-specific permease.

All these enzymes share the same catalytic mechanism: they bind PEP and transfer the phosphoryl group from it to a histidine residue. The sequence around that residue is highly conserved and can be used as a signature pattern for these enzymes. As a second signature pattern a conserved region was selected in the C-terminal part of the PEP-utilizing enzymes. The biological significance of this region is not yet known.

Consensus pattern: G-[GA]-x-[TN]-x-H-[STA]-[STAV]-[LIVM](2)-[STAV]-[RG] [H is phosphorylated]

-Consensus pattern: [DEQSK]-x-[LIVMF]-S-[LIVMF]-G-[ST]-N-D-[LIVM]-x-Q-[LIVMFYGT]-[STALIV]-[LIVMF]-[GAS]-x(2)-R

[1] Reizer J., Hoischen C., Reizer A., Pham T.N., Saier M.H. Jr. Protein Sci. 2:506-521(1993).

[2] Reizer J., Reizer A., Merrick M.J., Plunkett G. III, Rose D.J., Saier M.H. Jr. Gene 181:103-108(1996).

[3] Pocalyko D.J., Carroll L.J., Martin B.M., Babbitt P.C., Dunaway-Mariano D. Biochemistry 29:10757-10765(1990).

[4] Niersbach M., Kreuzaler F., Geerse R.H., Postma P., Hirsch H.J. Mol. Gen. Genet. 232:332-336(1992).

402. (PEPCK ATP) Phosphoenolpyruvate carboxykinase (ATP) signature
Phosphoenolpyruvate carboxykinase (ATP) (EC 4.1.1.49) (PEPCK) [1] catalyzes the formation of phosphoenolpyruvate by decarboxylation of oxaloacetate while hydrolyzing ATP, a rate limiting step in gluconeogenesis (the biosynthesis of glucose).

The sequence of this enzyme has been obtained from Escherichia coli, yeast,

and *Trypanosoma brucei*; these three sequences are evolutionary related and share many regions of similarity. As a signature pattern a highly conserved region was selected that contains four acidic residues and which is located in the central part of the enzyme. The beginning of the pattern is located about

5 10 residues to the C-terminus of an ATP-binding motif 'A' (P-loop) (see <PDOC00017>) and is also part of the ATP-binding domain [2].

Consensus pattern: L-I-G-D-D-E-H-x-W-x-[DE]-x-G-[IV]-x-N

-Note: phosphoenolpyruvate carboxykinase (GTP) (EC 4.1.1.32) an enzyme that catalyzes

10 the same reaction, but using GTP instead of ATP, is not related to the above enzyme (see <PDOC00421>).

[1] Medina V., Pontarollo R., Glaeske D., Tabel H., Goldie H. J. Bacteriol. 172:7151-7156(1990).

15 [2] Matte A., Goldie H., Sweet R.M., Delbaere L.T.J. J. Mol. Biol. 256:126-143(1996).

403. (Pepcase) Phosphoenolpyruvate carboxylase active sites. Phosphoenolpyruvate carboxylase (EC 4.1.1.31) (PEPcase) catalyzes the irreversible beta-carboxylation of phosphoenolpyruvate by bicarbonate to yield oxaloacetate and phosphate. The enzyme is found in all plants and in a variety of microorganisms. A histidine [1] and a lysine [2] have been implicated in the catalytic mechanism of this enzyme; the regions around these active site residues are highly conserved in PEPcase from various plants, bacteria and cyanobacteria and can be used as a signature patterns for this type of enzyme.

25 Consensus pattern: [VT]-x-T-A-H-P-T-[EQ]-x(2)-R-[KRH] [H is an active site residue]-

Consensus pattern: [IV]-M-[LIVM]-G-Y-S-D-S-x-K-D-[STAG]-G [K is an active site residue]-

30 [1] Terada K., Izui K. Eur. J. Biochem. 202:797-803(1991).[2] Jiao J.-A., Podesta F.E., Chollet R., O'Leary M.H., Andreo C.S. Biochim. Biophys. Acta 1041:291-295(1990).

404. PET112 family signature

The following proteins from eukaryotes, prokaryotes and archaeobacteria belong to the same family:

- Yeast mitochondrial protein PET112 [1], which plays an unknown role in the expression of mitochondrial genes, probably at the level of translation.
- *Aspergillus nidulans* mitochondrial protein nempA.
- *Bacillus subtilis* hypothetical protein yzdD.
- *Moraxella catarrhalis* hypothetical protein in bloR-1 3'region.
- *Mycoplasma genitalium* hypothetical protein MG100.
- *Methanococcus jannaschii* hypothetical proteins MJ0019 and MJ0160.

The size of these proteins range from 419 to 630 amino acids. As a signature pattern, a conserved region located in the N-terminal section was selected.

Consensus pattern: [DN]-x-[DN]-R-x(3)-P-L-[LIV]-E-[LIV]-x-[ST]-x-P

[1] Mulero J.J., Rosenthal J.K., Fox T.D. Curr. Genet. 25:299-304(1994).

405. (PFK) Phosphofructokinase signature

Phosphofructokinase (EC 2.7.1.11) (PFK) [1,2] is a key regulatory enzyme in the glycolytic pathway. It catalyzes the phosphorylation by ATP of fructose 6-phosphate to fructose 1,6-bisphosphate. In bacteria PFK is a tetramer of identical 36 Kd subunits. In mammals it is a tetramer of 80 Kd subunits. Each 80 Kd subunit consist of two homologous domains which are highly related to the bacterial 36 Kd subunits. In Human there are three, tissue-specific, types of PFK isozymes: PFKM (muscle), PFKL (liver), and PFKP (platelet). In yeast PFK is an octamer composed of four 100 Kd alpha chains (gene PFK1) and four 100 Kd beta chains (gene PFK2); like the mammalian 80 Kd subunits, the yeast 100 Kd subunits are composed of two homologous domains.

As a signature pattern for PFK a region that contains three basic residues involved in fructose-6-phosphate binding was selected.

Consensus pattern: [RK]-x(4)-G-H-x-Q-[QR]-G-G-x(5)-D-R [The R/K, the H and the Q/R are involved in fructose-6-P binding]

-Note: Escherichia coli has two phosphofructokinase isozymes which are encoded by genes pfkA (major) and pfkB (minor). The pfkB isozyme is not evolutionary related to other prokaryotic or eukaryotic PFK's (see <PDOC00504>).

[1] Poorman R.A., Randolph A., Kemp R.G., Heinrikson R.L. Nature 309:467-469(1984).
[2] Heinisch J., Ritzel R.G., von Borstel R.C., Aguilera A., Rodicio R., Zimmermann F.K. Gene 78:309-321(1989).

406. (PGAM) Phosphoglycerate mutase family phosphohistidine signature
Phosphoglycerate mutase (EC 5.4.2.1) (PGAM) and bisphosphoglycerate mutase (EC 5.4.2.4) (BPGM) are structurally related enzymes which catalyze reactions involving the transfer of phospho groups between the three carbon atoms of phosphoglycerate [1,2]. Both enzymes can catalyze three different reactions, although in different proportions:

- The isomerization of 2-phosphoglycerate (2-PGA) to 3-phosphoglycerate (3-PGA) with 2,3-diphosphoglycerate (2,3-DPG) as the primer of the reaction.
- The synthesis of 2,3-DPG from 1,3-DPG with 3-PGA as a primer.
- The degradation of 2,3-DPG to 3-PGA (phosphatase EC 3.1.3.13 activity).

In mammals, PGAM is a dimeric protein. There are two isoforms of PGAM: the M (muscle) and B (brain) forms. In yeast, PGAM is a tetrameric protein. BPGM is a dimeric protein and is found mainly in erythrocytes where it plays a major role in regulating hemoglobin oxygen affinity as a consequence of controlling 2,3-DPG concentration.

The catalytic mechanism of both PGAM and BPGM involves the formation of a phosphohistidine intermediate [3].

The bifunctional enzyme 6-phosphofructo-2-kinase / fructose-2,6-bisphosphatase (EC 2.7.1.105 and EC 3.1.3.46) (PF2K) [4] catalyzes both the synthesis and the degradation of fructose-2,6-bisphosphate. PF2K is an important enzyme in the regulation of hepatic carbohydrate metabolism. Like PGAM/BPGM, the fructose-2,6-bisphosphatase reaction involves a phosphohistidine intermediate and the

phosphatase domain of PF2K is structurally related to PGAM/BPGM.

The bacterial enzyme alpha-ribazole-5'-phosphate phosphatase (gene cobC) which is involved in cobalamin biosynthesis also belongs to this family [5].

A signature pattern was built around the phosphohistidine residue.

5

Consensus pattern: [LIVM]-x-R-H-G-[EQ]-x(3)-N [H is the phosphohistidine residue]

-Note: some organisms harbor a form of PGAM independent of 2,3-DPG, this enzyme is not related to the family described above [6].

10 [1] Le Boulch P., Joulin V., Garel M.-C., Rosa J., Cohen-Solal M. Biochem. Biophys. Res. Commun. 156:874-881(1988).

[2] White M.F., Fothergill-Gilmore L.A. FEBS Lett. 229:383-387(1988).

[3] Rose Z.B. Meth. Enzymol. 87:43-51(1982).

[4] Bazan J.F., Fletterick R.J., Pilgis S.J. Proc. Natl. Acad. Sci. U.S.A. 86:9642-9646(1989).

[5] O'Toole G.A., Trzebiatowski J.R., Escalante-Semerena J.C. J. Biol. Chem. 269:26503-26511(1994).

[6] Grana X., De Lecea L., El-Maghrabi M.R., Urena J.M., Caellas C., Carreras J., Puigdomenech P., Pilgis S.J., Climent F. J. Biol. Chem. 267:12797-12803(1992).

20

407. (PGI) Phosphoglucose isomerase signatures

Phosphoglucose isomerase (EC 5.3.1.9) (PGI) [1,2] is a dimeric enzyme that catalyzes the reversible isomerization of glucose-6-phosphate and fructose-6-phosphate. PGI is involved in different pathways: in most higher organisms it is involved in glycolysis; in mammals it is involved in gluconeogenesis; in plants in carbohydrate biosynthesis; in some bacteria it provides a gateway for fructose into the Entner-Doudouroff pathway. PGI has been shown [3] to be identical to neuroleukin, a neurotrophic factor which supports the survival of various types of neurons.

30

The sequence of PGI from many species ranging from bacteria to mammals is available and has been shown to be highly conserved. As signature patterns for this enzyme two conserved regions were selected, the first region is located in

the central section of PGI, while the second one is located in its C-terminal section.

Consensus pattern: [DENS]-x-[LIVM]-G-G-R-[FY]-S-[LIVMT]-x-[STA]-[PSAC]-
[LIVMA]-G

-Consensus pattern: [GS]-x-[LIVM]-[LIVMFYW]-x(4)-[FY]-[DN]-Q-x-G-V-E-x(2)-K

[1] Achari A., Marshall S.E., Muirhewad H., Palmieri R.H., Noltmann E.A. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 293:145-157(1981).

[2] Smith M.W., Doolittle R.F. J. Mol. Evol. 34:544-545(1992).

[3] Faik P., Walker J.I.H., Redmill A.A.M., Morgan M.J. Nature 332:455-456(1988).

408. (PGK) Phosphoglycerate kinase signature

Phosphoglycerate kinase (EC 2.7.2.3) (PGK) [1] catalyzes the second step in the second phase of glycolysis, the reversible conversion of 1,3-diphosphoglycerate to 3-phosphoglycerate with generation of one molecule of ATP. PGK is found in all living organisms and its sequence has been highly conserved throughout evolution. It is a two-domain protein; each domain is composed of six repeats of an alpha/beta structural motif. As a signature pattern for PGK's, a conserved region in the N-terminal region was selected.

Consensus pattern: [KRHGTCVN]-[VT]-[LIVMF]-[LIVMC]-R-x-D-x-N-[SACV]-P

[1] Watson H.C., Littlechild J.A. Biochem. Soc. Trans. 18:187-190(1990).

409. (PGM PMM) Phosphoglucomutase and phosphomannomutase phosphoserine signature

- Phosphoglucomutase (EC 5.4.2.2) (PGM). PGM is an enzyme responsible for the conversion of D-glucose 1-phosphate into D-glucose 6-phosphate. PGM participates in both the breakdown and synthesis of glucose [1].

- Phosphomannomutase (EC 5.4.2.8) (PMM). PMM is an enzyme responsible for the conversion of D-mannose 1-phosphate into D-mannose 6-phosphate. PMM is required for different biosynthetic pathways in bacteria. For example, in

enterobacteria such as *Escherichia coli* there are two different genes coding for this enzyme: *rfbK* which is involved in the synthesis of the O antigen of lipopolysaccharide and *cpsG* which is required for the synthesis of the M antigen capsular polysaccharide [2]. In *Pseudomonas aeruginosa* PMM (gene *algC*) is involved in the biosynthesis of the alginate layer [3] and in *Xanthomonas campestris* (gene *xanA*) it is involved in the biosynthesis of xanthan [4]. In *Rhizobium* strain ngr234 (gene *noeK*) it is involved in the biosynthesis of the nod factor.

- Phosphoacetylglucosamine mutase (EC 5.4.2.3) which converts N-acetyl-D-glucosamine 1-phosphate into the 6-phosphate isomer.

The catalytic mechanism of both PGM and PMM involves the formation of a phosphoserine intermediate [1]. The sequence around the serine residue is well conserved and can be used as a signature pattern.

In addition to PGM and PMM there are at least three uncharacterized proteins that belong to this family [5,6]:

- Urease operon protein *ureC* from *Helicobacter pylori*.
- *Escherichia coli* protein *mrsA*.
- *Paramecium tetraurelia* *parafusin*, a phosphoglycoprotein involved in exocytosis.
- A *Methanococcus vannielii* hypothetical protein in the 3' region of the gene for ribosomal protein S10.

Consensus pattern: [GSA]-[LIVM]-x-[LIVM]-[ST]-[PGA]-S-H-x-P-x(4)-[GNHE] [S is the phosphoserine residue]

-Note: PMM from fungi do not belong to this family.

[1] Dai J.B., Liu Y., Ray W.J. Jr., Konno M. J. Biol. Chem. 267:6322-6337(1992).

[2] Stevenson G., Lee S.J., Romana L.K., Reeves P.R. Mol. Gen. Genet. 227:173-180(1991).

[3] Zielinski N.A., Chakrabarty A.M., Berry A. J. Biol. Chem. 266:9754-9763(1991).

[4] Koeplin R., Arnold W., Hoette B., Simon R., Wang G., Puehler A. J. Bacteriol. 174:191-199(1992).

[5] Bairoch A. Unpublished observations (1993).

[6] Subramanian S.V., Wyroba E., Andersen A.P., Satir B.H. Proc. Natl. Acad. Sci. U.S.A. 91:9832-9836(1994).

5 410. PH domain profile

The 'pleckstrin homology' (PH) domain is a domain of about 100 residues that occurs in a wide range of proteins involved in intracellular signaling or as constituents of the cytoskeleton [1 to 7].

10 The function of this domain is not clear, several putative functions have been suggested: - binding to the beta/gamma subunit of heterotrimeric G proteins,

- binding to lipids, e.g. phosphatidylinositol-4,5-bisphosphate,

- binding to phosphorylated Ser/Thr residues,

- attachment to membranes by an unknown mechanism.

15 It is possible that different PH domains have totally different ligand requirements.

The 3D structure of several PH domains has been determined [8]. All known cases have a common structure consisting of two perpendicular anti-parallel beta sheets, followed by a C-terminal amphipathic helix. The loops connecting the beta-strands differ greatly in length, making the PH domain relatively
20 difficult to detect. There are no totally invariant residues within the PH domain.

Proteins reported to contain one more PH domains belong to the following families:

- Pleckstrin, the protein where this domain was first detected, is the major
25 substrate of protein kinase C in platelets. Pleckstrin is one of the rare proteins to contains two PH domains.
- Ser/Thr protein kinases such as the Act/Rac family, the beta-adrenergic receptor kinases, the mu isoform of PKC and the trypanosomal NrKA family.
- Tyrosine protein kinases belonging to the Btk/Itk/Tec subfamily.
- 30 - Insulin Receptor Substrate 1 (IRS-1).
- Regulators of small G-proteins like guanine nucleotide releasing factor GNRP (Ras-GRF) (which contains 2 PH domains), guanine nucleotide exchange proteins like vav, dbl, SoS and yeast CDC24, GTPase activating proteins

like rasGAP and BEM2/IPL2, and the human break point cluster protein bcr.

- Cytoskeletal proteins such as dynamin (see <PDOC00362>), *Caenorhabditis elegans* kinesin-like protein unc-104 (see <PDOC00343>), spectrin beta-chain, syntrophin (2 PH domains) and yeast nuclear migration protein NUM1.
- 5 - Mammalian phosphatidylinositol-specific phospholipase C (PI-PLC) (see <PDOC50007>) isoforms gamma and delta. Isoform gamma contains two PH domains, the second one is split into two parts separated by about 400 residues.
- Oxysterol binding proteins OSBP, yeast OSH1 and YHR073w.
- Mouse protein citron, a putative rho/rac effector that binds to the GTP-bound forms of rho and rac,
- 10 - Several yeast proteins involved in cell cycle regulation and bud formation like BEM2, BEM3, BUD4 and the BEM1-binding proteins BOI2 (BEB1) and BOI1 (BOB1).
- *Caenorhabditis elegans* protein MIG-10.
- *Caenorhabditis elegans* hypothetical proteins C04D8.1, K06H7.4 and ZK632.12.
- 5 - Yeast hypothetical proteins YBR129c and YHR155w.

The profile for the PH domain, which has been developed by Toby Gibson at the EMBL, covers the total length of domain. Several proteins contain large insertions in the PH domain and are thus difficult to detect with this profile. In some of these cases, the profile will align only to one half of the PH domain.

-Sequences known to belong to this class detected by the pattern: ALL. But it should be noted that while all sequences containing PH domains are detected, not all PH domains are. Some of the split domains lie below the cutoff threshold.

- 25 [1] Mayer B.J., Ren R., Clark K.L., Baltimore D. Cell 73:629-630(1993).
- [2] Haslam R.J., Koide H.B., Hemmings B.A. Nature 363:309-310(1993).
- [3] Musacchio A., Gibson T.J., Rice P., Thompson J., Saraste M. Trends Biochem. Sci. 18:343-348(1993).
- [4] Gibson T.J., Hyvonen M., Musacchio A., Saraste M., Birney E. Trends Biochem. Sci. 19:349-353(1994).
- 30 [5] Pawson T. Nature 373:573-580(1995).
- [6] Ingle E., Hemmings B.A. J. Cell. Biochem. 56:436-443(1994).
- [7] Saraste M., Hyvonen M. Curr. Opin. Struct. Biol. 5:403-408(1995).
- [8] Riddihough G.

Nat. Struct. Biol. 1:755-757(1994).

411. PHD-finger

5 [1]

Medline: 95216093

The PHD finger: implications for chromatin-mediated transcriptional regulation.

Aasland R, Gibson TJ, Stewart AF;

10 Trends Biochem Sci 1995;20:56-59.

Number of members: 181

412. (PI-PLC-X) Phosphatidylinositol-specific phospholipase C profiles

15 Phosphatidylinositol-specific phospholipase C (EC 3.1.4.11), an eukaryotic intracellular enzyme, plays an important role in signal transduction processes [1]. It catalyzes the hydrolysis of 1-phosphatidyl-D-myo-inositol-3,4,5-triphosphate into the second messenger molecules diacylglycerol and inositol-1,4,5-triphosphate. This catalytic process is tightly regulated by reversible phosphorylation and binding of regulatory proteins [2 to 4].

20 In mammals, there are at least 6 different isoforms of PI-PLC, they differ in their domain structure, their regulation, and their tissue distribution. Lower eukaryotes also possess multiple isoforms of PI-PLC.

25 All eukaryotic PI-PLCs contain two regions of homology, sometimes referred to as 'X-box' and 'Y-box'. The order of these two regions is always the same (NH2-X-Y-COOH), but the spacing is variable. In most isoforms, the distance between these two regions is only 50-100 residues but in the gamma isoforms one PH domain, two SH2 domains, and one SH3 domain are inserted between the two PLC-specific domains. The two conserved regions have been shown to be 30 important for the catalytic activity. At the C-terminal of the Y-box, there is a C2 domain (see <PDOC00380>) possibly involved in Ca-dependent membrane attachment.

Profile analysis shows that sequences with significant similarity

to the X-box domain occur also in prokaryotic and trypanosome PI-specific phospholipases C. Apart from this region, the prokaryotic enzymes show no similarity to their eukaryotic counterparts.

Two profiles were developed, one covering the X-box, the other the Y-box.

- 5 [1] Meldrum E., Parker P.J., Carozzi A.
Biochim. Biophys. Acta 1092:49-71(1991).[2] Rhee S.G., Choi K.D.
Adv. Second Messenger Phosphoprotein Res. 26:35-61(1992).
[3] Rhee S.G., Choi K.D. J. Biol. Chem. 267:12393-12396(1992).
[4] Sternweis P.C., Smrcka A.V. Trends Biochem. Sci. 17:502-506(1992).

10

413. (PI-PLC-Y) Phosphatidylinositol-specific phospholipase C profiles

Phosphatidylinositol-specific phospholipase C (EC 3.1.4.11), an eukaryotic intracellular enzyme, plays an important role in signal transduction processes [1]. It catalyzes the hydrolysis of 1-phosphatidyl-D-myo-inositol-3,4,5-triphosphate into the second messenger molecules diacylglycerol and inositol-1,4,5-triphosphate. This catalytic process is tightly regulated by reversible phosphorylation and binding of regulatory proteins [2 to 4].

In mammals, there are at least 6 different isoforms of PI-PLC, they differ in their domain structure, their regulation, and their tissue distribution. Lower eukaryotes also possess multiple isoforms of PI-PLC.

All eukaryotic PI-PLCs contain two regions of homology, sometimes referred to as 'X-box' and 'Y-box'. The order of these two regions is always the same (NH₂-X-Y-COOH), but the spacing is variable. In most isoforms, the distance between these two regions is only 50-100 residues but in the gamma isoforms one PH domain, two SH2 domains, and one SH3 domain are inserted between the two PLC-specific domains. The two conserved regions have been shown to be important for the catalytic activity. At the C-terminal of the Y-box, there is a C2 domain (see <PDOC00380>) possibly involved in Ca-dependent membrane attachment.

Profile analysis shows that sequences with significant similarity to the X-box domain occur also in prokaryotic and trypanosome PI-specific phospholipases C. Apart from this region, the prokaryotic enzymes show no

similarity to their eukaryotic counterparts.

Two profiles were developed, one covering the X-box, the other the Y-box.

[1] Meldrum E., Parker P.J., Carozzi A.

Biochim. Biophys. Acta 1092:49-71(1991).[2] Rhee S.G., Choi K.D.

5 Adv. Second Messenger Phosphoprotein Res. 26:35-61(1992).

[3] Rhee S.G., Choi K.D. J. Biol. Chem. 267:12393-12396(1992).

[4] Sternweis P.C., Smrcka A.V. Trends Biochem. Sci. 17:502-506(1992).

10 414. (PK) Pyruvate kinase active site signature

Pyruvate kinase (EC 2.7.1.40) (PK) [1] catalyzes the final step in glycolysis, the conversion of phosphoenolpyruvate to pyruvate with the concomitant phosphorylation of ADP to ATP. PK requires both magnesium and potassium ions for its activity. PK is found in all living organisms. In vertebrates there are four, tissues specific, isozymes: L (liver), R (red cells), M1 (muscle, heart, and brain), and M2 (early fetal tissues). In Escherichia coli there are two isozymes: PK-I (gene pykF) and PK-II (gene pykA). All PK isozymes seem to be tetramers of identical subunits of about 500 amino acid residues.

As a signature pattern for PK a conserved region was selected that includes a lysine residue which seems to be the acid/base catalyst responsible for the interconversion of pyruvate and enolpyruvate, and a glutamic acid residue implicated in the binding of the magnesium ion.

Consensus pattern: [LIVAC]-x-[LIVM](2)-[SAPCV]-K-[LIV]-E-[NKRST]-x-[DEQHS]-

25 [GSTA]-[LIVM] [K is the active site residue] [E is a magnesium ligand]

[1] Muirhead H. Biochem. Soc. Trans. 18:193-196(1990).

30 415. (PLDc) Phospholipase D. Active site motif

Phosphatidylcholine-hydrolyzing phospholipase D (PLD) isoforms are activated by ADP-ribosylation factors (ARFs). PLD produces phosphatidic acid from phosphatidylcholine, which may be essential for the formation

of certain types of transport vesicles or may be constitutive vesicular transport to signal transduction pathways.

PC-hydrolyzing PLD is a homologue of cardiolipin synthase, phosphatidylserine synthase, bacterial PLDs, and viral proteins.

5 Each of these appears to possess a domain duplication which is apparent by the presence of two motifs containing well-conserved histidine, lysine, and/or asparagine residues which may contribute to the active site. aspartic acid. An *E. coli* endonuclease (nuc) and similar proteins appear to be PLD homologues but possess only one of these motifs.

10 The profile contained here represents only the putative active site regions, since an accurate multiple alignment of the repeat units has not been achieved.

Number of members: 139

[1]

5 Medline: 96303814

A novel family of phospholipase D homologues that includes phospholipid synthases and putative endonucleases: identification of duplicated repeats and potential active site residues.

20 Ponting CP, Kerr ID;
Protein Sci 1996;5:914-922.

[2]Medline: 96334293

A duplicated catalytic motif in a new superfamily of phosphohydrolases and phospholipid synthases that includes poxvirus envelope proteins.

25 Koonin EV;
Trends Biochem Sci 1996;21:242-243.

[3]Medline: 94327597

Cloning and expression of phosphatidylcholine-hydrolyzing phospholipase D from *Ricinus communis* L.

30 Wang X, Xu L, Zheng L;
J Biol Chem 1994;269:20312-20317.

[4]Medline: 97386825

Regulation of eukaryotic phosphatidylinositol-specific
phospholipase C and phospholipase D.

Singer WD, Brown HA, Sternweis PC;
Annu Rev Biochem 1997;66:475-509.

5

416. (PMI type1) Phosphomannose isomerase type I signatures

Phosphomannose isomerase (EC 5.3.1.8) (PMI) [1,2] is the enzyme that catalyzes
the interconversion of mannose-6-phosphate and fructose-6-phosphate. In
eukaryotes, it is involved in the synthesis of GDP-mannose which is a
constituent of N- and O-linked glycans as well as GPI anchors. In prokaryotes,
it is involved in a variety of pathways including capsular polysaccharide
biosynthesis and D-mannose metabolism.

10

Three classes of PMI have been defined on the basis of sequence similarities
[1]. The first class comprises all known eukaryotic PMI as well as the enzyme
encoded by the *manA* gene in enterobacteria such as *Escherichia coli*. Class I
PMI's are proteins of about 42 to 50 Kd which bind a zinc ion essential for
their activity.

15

As signature patterns for class I PMI, two conserved regions were selected. The
first one is located in the N-terminal section of these proteins, the second
in the C-terminal half. Both patterns contain a residue involved [3] in the
binding of the zinc ion.

20

Consensus pattern: Y-x-D-x-N-H-K-P-E [E is a zinc ligand]

25

-Consensus pattern: H-A-Y-[LIVM]-x-G-x(2)-[LIVM]-E-x-M-A-x-S-D-N-x-[LIVM]-R-A-
G-x-T-P-K [H is a zinc ligand]

[1] Proudfoot A.E.I., Turcatti G., Wells T.N.C., Payton M.A., Smith D.J. Eur. J. Biochem.
219:415-423(1994).

30

[2] Coulin F., Magnenat E., Proudfoot A.E.I., Payton M.A., Scully P., Wells T.N.C.
Biochemistry 32:14139-14144(1993).

[3] Cleasby A., Wonacott A., Skarzynski T., Hubbard R.E., Davies G.J., Proudfoot A.E.I.,
Bernard A.R., Payton M.A., Wells T.N.C. Nat. Struct. Biol. 3:470-479(1996).

417. (PNP UDP 1) Purine and other phosphorylases family 1 signature

The following phosphorylases belongs to the same family:

- Purine nucleoside phosphorylase (EC 2.4.2.1) (PNP) from most bacteria (gene deoD). This enzyme catalyzes the cleavage of guanosine or inosine to respective bases and sugar-1-phosphate molecules [1].
- Uridine phosphorylase (EC 2.4.2.3) (UdRPase) from bacteria (gene udp) and mammals. Catalyzes the cleavage of uridine into uracil and ribose-1-phosphate. The products of the reaction are used either as carbon and energy sources or in the rescue of pyrimidine bases for nucleotide synthesis [2].
- 5'-methylthioadenosine phosphorylase (EC 2.4.2.28) (MTA phosphorylase) from *Sulfolobus solfataricus* [3].

As a signature pattern, a conserved region was selected in the central part of these enzymes.

Consensus pattern: [GST]-x-G-[LIVM]-G-x-[PA]-S-x-[GSTA]-I-x(3)-E-L

-Note: it should be noted that mammalian and some bacterial PNP as well as eukaryotic MTA phosphorylase belong to a different family of phosphorylases (see <PDOC00954>).

[1] Takehara M., Ling F., Izawa S., Inoue Y., Kimura A. Biosci. Biotechnol. Biochem. 59:1987-1990(1995).

[2] Watanabe S.-I., Hino A., Wada K., Eliason J.F., Uchida T. J. Biol. Chem. 270:12191-12196(1995).

[3] Cacciapuoti G., Porcelli M., Bertoldo C., De Rosa M., Zappia V. J. Biol. Chem. 269:24762-24769(1994).

418. (PP2C) Protein phosphatase 2C signature

Protein phosphatase 2C (PP2C) is one of the four major classes of mammalian serine/threonine specific protein phosphatases (EC 3.1.3.16). PP2C [1] is a monomeric enzyme of about 42 Kd which shows broad substrate specificity and

is dependent on divalent cations (mainly manganese and magnesium) for its activity. Its exact physiological role is still unclear. Three isozymes are currently known in mammals: PP2C- α , - β and - γ . In yeast, there are at least four PP2C homologs: phosphatase PTC1 [2] which has weak tyrosine phosphatase activity in addition to its activity on serines, phosphatases PTC2 and PTC3, and hypothetical protein YBR125c. Isozymes of PP2C are also known from *Arabidopsis thaliana* (ABI1, PPH1), *Caenorhabditis elegans* (FEM-2, F42G9.1, T23F11.1), *Leishmania chagasi* and *Paramecium tetraurelia*. In *Arabidopsis thaliana*, the kinase associated protein phosphatase (KAPP) [3] is an enzyme that dephosphorylates the Ser/Thr receptor-like kinase RLK5 and which contains a C-terminal PP2C domain.

PP2C does not seem to be evolutionary related to the main family of serine/threonine phosphatases: PP1, PP2A and PP2B. However, it is significantly similar to the catalytic subunit of pyruvate dehydrogenase phosphatase (EC 3.1.3.43) (PDPC) [4], which catalyzes dephosphorylation and concomitant reactivation of the α subunit of the E1 component of the pyruvate dehydrogenase complex. PDPC is a mitochondrial enzyme and, like PP2C, is magnesium-dependent.

As a signature pattern, the best conserved region was selected which is located in the N-terminal part and contains a perfectly conserved tripeptide. This region includes a conserved aspartate residue involved in divalent cation binding [5].

Consensus pattern: [LIVMFY]-[LIVMFYA]-[GSAC]-[LIVM]-[FYC]-D-G-H-[GAV]

-Note: PP2C belongs [6] to a superfamily which also includes bacterial proteins such as *Bacillus spoIIIE*, *rsbU* and *rsbW*, *Synechocystis* PCC 6803 *icfG* as well as a domain in fungal adenylate cyclases.

[1] Wenk J., Trompeter H.-I., Pettrich K.-G., Cohen P.T.W., Campbell D.G., Mieskes G. FEBS Lett. 297:135-138(1992).

[2] Maeda T., Tsai A.Y.M., Saito H. Mol. Cell. Biol. 13:5408-5417(1993).

[3] Stone J.M., Collinge M.A., Smith R.D., Horn M.A., Walker J.C. Science 266:793-795(1994).

[4] Lawson J.E., Niu X.-D., Browning K.S., Trong H.L., Yan J., Reed L.J. *Biochemistry* 32:8987-8993(1993).

[5] Das A.K., Helps N.R., Cohen P.T.W., Barford D. *EMBO J.* 24:6798-6809(1996).

[6] Bork P., Brown N.P., Hegyi H., Schultz J. *Protein Sci.* 5:1421-1425(1996).

5

419. (PPTA) Protein prenyltransferases alpha subunit repeat signature

Protein prenyltransferases catalyze the transfer of an isoprenyl moiety to a cysteine four residues from the C-terminus of several proteins. They are heterodimeric enzymes consisting of alpha and beta subunits. The alpha subunit is thought to participate in a stable complex with the isoprenyl substrate; the beta subunit binds the peptide substrate. Distinct protein prenyltransferases might share a common alpha subunit. Both the alpha and beta subunit show repetitive sequence motifs [1]. These repeats have distinct structural and functional implications and are unrelated to each other. Known protein prenyltransferase alpha subunits are:

- Mammalian protein farnesyltransferase alpha subunit.
- Yeast protein RAM2, a protein farnesyltransferase alpha subunit.
- Yeast protein BET4, a protein geranylgeranyltransferase alpha subunit.

The conserved domain of the alpha subunit consists of about 34 amino acids and is repeated five times. It contains an invariant tryptophan possibly involved in heterodimerization with the conserved phenylalanines in the repeated domains of the beta subunits, via hydrophobic bonds. The signature pattern for this domain is centered on the invariant tryptophan.

25

Consensus pattern: [PSIAV]-x-[NDFV]-[NEQIY]-x-[LIVMAGP]-W-[NQSTHF]-[FYHQ]-[LIVMR]

[1] Boguski M.S., Murray A.W., Powers S. *New Biol.* 4:408-411(1992).

30

420. (PR55) Protein phosphatase 2A regulatory subunit PR55 signatures

Protein phosphatase 2A (PP2A) is a serine/threonine phosphatase involved in

many aspects of cellular function including the regulation of metabolic enzymes and proteins involved in signal transduction. PP2A is a trimeric enzyme that consists of a core composed of a catalytic subunit associated with a 65 Kd regulatory subunit (PR65), also called subunit A; this complex then associates with a third variable subunit (subunit B), which confers distinct properties to the holoenzyme [1]. One of the forms of the variable subunit is a 55 Kd protein (PR55) which is highly conserved in mammals - where three isoforms are known to exist -, *Drosophila* and yeast (gene CDC55). This subunit could perform a substrate recognition function or be responsible for targeting the enzyme complex to the appropriate subcellular compartment.

As signature patterns, two perfectly conserved sequences of 15 residues were selected; one located in the N-terminal region, the other in the center of the protein.

Consensus pattern: E-F-D-Y-L-K-S-L-E-I-E-E-K-I-N

Consensus pattern: N-[AG]-H-[TA]-Y-H-I-N-S-I-S-[LIVM]-N-S-D

[1] Mayer-Jaekel R., Hemmings B.A. Trends Cell Biol. 4:287-291(1994).

421. N-(5'phosphoribosyl)anthranilate (PRA) isomerase

[1] Wilmanns M, Priestle JP, Niermann T, Jansonius JN;
J Mol Biol 1992;223:477-507.

422. (PRK) Phosphoribulokinase signature

Phosphoribulokinase (EC 2.7.1.19) (PRK) [1,2] is one of the enzymes specific to the Calvin's reductive pentose phosphate cycle which is the major route by which carbon dioxide is assimilated and reduced by autotrophic organisms. PRK catalyzes the ATP-dependent phosphorylation of ribulose 5-phosphate into ribulose 1,5-bisphosphate which is the substrate for RubisCO.

PRK's of diverse origins show different properties with respect to the size of the protein, the subunit structure, or the enzymatic regulation. However an

alignment of the sequences of PRK from plants, algae, photosynthetic and chemoautotrophic bacteria shows that there are a few regions of sequence similarity. As a signature pattern one of these regions was selected.

5 Consensus pattern: K-[LIVM]-x-R-D-x(3)-R-G-x-[ST]-x-E

[1] Kossmann J., Klintworth R., Bowien B. Gene 85:247-252(1989).

[2] Gibson J.L., Chen J.-H., Tower P.A., Tabita F.R. Biochemistry 29:8085-8093(1990).

10

423. (PRPP synt) Phosphoribosyl pyrophosphate synthetase signature

Phosphoribosyl pyrophosphate synthetase (EC 2.7.6.1) (PRPP synthetase) catalyzes the formation of PRPP from ATP and ribose 5-phosphate. PRPP is then used in various biosynthetic pathways, as for example in the formation of purines, pyrimidines, histidine and tryptophan. PRPP synthetase requires inorganic phosphate and magnesium ions for its stability and activity.

In mammals, three isozymes of PRPP synthetase are found; in yeast there are at least four isozymes.

As a signature pattern for this enzyme, a very conserved region was selected that has been suggested to be involved in binding divalent cations [1]. This region contains two conserved aspartic acid residues as well as a histidine, which are all potential ligands for a cation such as magnesium.

Consensus pattern: D-[LI]-H-[SA]-x-Q-[IMST]-[QM]-G-[FY]-F-x(2)-P-[LIVMFC]-D

25

[1] Bower S.G., Harlow K.W., Switzer R.L., Hoven-Jensen B. J. Biol. Chem. 264:10287-10291(1989).

30 424. (PRTP) Herpesvirus processing and transport protein

The members of this family are associate with capsid intermediates during packaging of the virus.

Number of members: 31

[1]

Medline: 98362148

Herpes simplex virus type 1 cleavage and packaging proteins
UL15 and UL28 are associated with B but not C capsids during
packaging. Yu D, Weller SK;
J Virol 1998;72:7428-7439.

425. Photosystem I psaG / psaK (PSI PSAK) proteins signature

Photosystem I (PSI) [1] is an integral membrane protein complex that uses light energy to mediate electron transfer from plastocyanin to ferredoxin. It is found in the chloroplasts of plants and cyanobacteria. PSI is composed of at least 14 different subunits, two of which PSI-G (gene psaG) and PSI-K (gene psaK) are small hydrophobic proteins of about 7 to 9 Kd and evolutionary related [2]. Both seem to contain two transmembrane regions. Cyanobacteria seem to encode only for PSI-K.

As a signature pattern, the best-conserved region was selected which seems to correspond to the second transmembrane region.

-Consensus pattern: [GT]-F-x-[LIVM]-x-[DEA]-x(2)-[GA]-x-[GTA]-[SA]-x-G-H-x-[LIVM]-[GA]

[1] Golbeck J.H. Biochim. Biophys. Acta 895:167-204(1987).

[2] Kjaerulff S., Andersen B., Nielsen V.S., Moller B.L., Okkels J.S. J. Biol. Chem. 268:18912-18916(1993).

426. PTR2 family proton/oligopeptide symporters signatures

A family of eukaryotic and prokaryotic proteins that seem to be mainly involved in the intake of small peptides with the concomitant uptake of a proton has been recently characterized [1,2]. Proteins that belong to this family are: - Fungal peptide transporter PTR2.

- Mammalian intestine proton-dependent oligopeptide transporter PeptT1.
- Mammalian kidney proton-dependent oligopeptide transporter PeptT2.

- *Drosophila opt1*.
- *Arabidopsis thaliana* peptide transporters PTR2-A and PTR2-B (also known as the histidine transporting protein NTR1).
- *Arabidopsis thaliana* proton-dependent nitrate/chlorate transporter CHL1.
- 5 - *Lactococcus* proton-dependent di- and tri-peptide transporter dtpT.
- *Caenorhabditis elegans* hypothetical protein C06G8.2.
- *Caenorhabditis elegans* hypothetical protein F56F4.5.
- *Caenorhabditis elegans* hypothetical protein K04E7.2.
- *Escherichia coli* hypothetical protein ybgH.
- 10 - *Escherichia coli* hypothetical protein ydgR.
- *Escherichia coli* hypothetical protein yhiP.
- *Escherichia coli* hypothetical protein yjdL.
- *Bacillus subtilis* hypothetical protein yclF.

These integral membrane proteins are predicted to comprise twelve transmembrane regions. As signature patterns, two of the best conserved regions were selected. The first is a region that includes the end of the second transmembrane region, a cytoplasmic loop as well as the third transmembrane region. The second pattern corresponds to the core of the fifth transmembrane region.

-Consensus pattern: [GA]-[GAS]-[LIVMFYWA]-[LIVM]-[GAS]-D-x-[LIVMFYWT]-[LIVMFYW]-G-x(3)-[TAV]-[IV]-x(3)-[GSTAV]-x-[LIVMF]-x(3)-[GA]
 -Consensus pattern: [FYT]-x(2)-[LMFY]-[FYV]-[LIVMFYWA]-x-[IVG]-N-[LIVMAG]-G-[GSA]-[LIMF]

- [1] Paulsen I.T., Skurray R.A. Trends Biochem. Sci. 19:404-404(1994).
- [2] Steiner H.-Y., Naider F., Becker J.M. Mol. Microbiol. 16:825-834(1995).

427. Pumilio-family RNA binding domains (aka PUM-HD, Pumilio homology domain)

Puf domains are necessary and sufficient for sequence specific RNA binding in fly Pumilio and worm FBF-1 and FBF-2. Both proteins

function as translational repressors in early embryonic development by binding sequences in the 3' UTR of target mRNAs (e.g. the nanos response element (NRE) in fly Hunchback mRNA, or the point mutation element (PME) in worm fem-3 mRNA). Other proteins that contain Puf domains are also plausible RNA binding proteins. JSN1_YEAST, for instance, appears to also contain a single RRM domain by HMM analysis.

Puf domains usually occur as a tandem repeat of 8 domains.

The Pfam model does not necessarily recognize all 8 domains in all sequences; some sequences appear to have 5 or 6 domains on initial analysis, but further analysis suggests the presence of additional divergent domains.

[1] Zhang B, Gallegos M, Puoti A, Durkin E, Fields S, Kimble J, Wickens MP. Nature 1997;390:477-484. [2] Zamore PD, Williamson JR, Lehmann R. RNA 1997;3:1421-1433.

428. PWWP domain. The PWWP domain is named after a conserved Pro-Trp-Trp-Pro motif. The function of the domain is currently unknown. Number of members: 19

[1] Medline: 98282232. WHSC1, a 90 kb SET domain-containing gene, expressed in early development and homologous to a Drosophila dysmorphia gene maps in the Wolf-Hirschhorn syndrome critical region and is fused to IgH in t(4;14) multiple myeloma. Stec I, Wright TJ, van Ommen GJB, de Boer PAJ, van Haeringen A, Moorman AFM, Altherr MR, den Dunnen JT; Hum Mol Genet 1998;7:1071-1082.

429. PX domain

Eukaryotic domain of unknown function present in phox proteins, PLD isoforms, a PI3K isoform.

Number of members: 71

[1]

Medline: 97084820

Novel domains in NADPH oxidase subunits, sorting nexins, and
PtdIns 3-kinases: binding partners of SH3 domains?

Ponting CP;

Protein Sci 1996;5:2353-2357.

5

430. ParA family ATPase

[1]

Medline: 91141297

10 A family of ATPases involved in active partitioning of
diverse bacterial plasmids.

Motallebi-Veshareh M, Rouch DA, Thomas CM;

Mol Microbiol 1990;4:1455-1463.

Number of members: 122

15

431. (Parvo coat) Parvovirus coat protein. 72 members.

20

432. Pectinesterase signatures

Pectinesterase (EC 3.1.1.11) (pectin methylesterase) catalyzes the hydrolysis
of pectin into pectate and methanol. In plants, it plays an important role in
cell wall metabolism during fruit ripening. In plant bacterial pathogens such
as *Erwinia carotovora* and in fungal pathogens such as *Aspergillus niger*,
pectinesterase is involved in maceration and soft-rotting of plant tissue.

25

Prokaryotic and eukaryotic pectinesterases share a few regions of sequence
similarity [1,2,3]. two of these regions were selected as signature patterns.

The first is based on a region in the N-terminal section of these enzymes; it
contains a conserved tyrosine which may play a role in the catalytic mechanism
[3]. The second pattern corresponds to the best conserved region, an
octapeptide located in the central part of these enzymes.

30

-Consensus pattern: [GSTNP]-x(6)-[FYVHR]-[IVN]-[KEP]-x-G-[STIVKRQ]-Y-
[DNQKRMV]-[EP]-x(3)-[LIMVA]

-Consensus pattern: [IV]-x-G-[STAD]-[LIVT]-D-[FYI]-[IV]-[FSN]-G

- 5 [1] Ray J., Knapp J., Grierson D., Bird C., Schuch W. Eur. J. Biochem. 174:119-124(1988).
[2] Plastow G.S. Mol. Microbiol. 2:247-254(1988).
[3] Markovic O., Joernvall H. Protein Sci. 1:1288-1292(1992).

10 433. Pentapeptide repeats (8 copies)

These repeats are found in many cyanobacterial proteins.

The repeats were first identified in hglK [1]. The function of these repeats is unknown.

The structure of this repeat has been predicted to be a beta-helix [2].

The repeat can be approximately described as A(D/N)LXX, where X can be any amino acid. Number of members: 75

[1]

Medline: 96062225

20 The hglK gene is required for localization of heterocyst-specific glycolipids in the cyanobacterium Anabaena sp. strain PCC 7120.

Black K, Buikema WJ, Haselkorn R;

J Bacteriol 1995;177:6440-6448.

25 [2]Medline: 98318059

Structure and distribution of pentapeptide repeats in bacteria.

Bateman A, Murzin A, Teichmann SA;

Protein Sci 1998;7:1477-1480.

30 [3]Medline: 98316713

Characterisation of an Arabidopsis cDNA encoding a thylakoid lumen protein related to a novel 'pentapeptide repeat' family of proteins.

Kieselbach T, Mant A, Robinson C, Schroder WP;
FEBS Lett 1998;428:241-244.

5 434. Polypeptide deformylase
[1]

Medline: 97002011

A new subclass of the zinc metalloproteases superfamily
revealed by the solution structure of peptide deformylase.

10 Meinnel T, Blanquet S, Dardel F;
J Mol Biol 1996;262:375-386.

[2]Medline: 98332750

Solution structure of nickel-peptide deformylase.

Dardel F, Ragusa S, Lazennec C, Blanquet S, Meinnel T;
J Mol Biol 1998;280:501-513.

Number of members: 21

435. Peptidyl-tRNA hydrolase signatures

20 Peptidyl-tRNA hydrolase (EC 3.1.1.29) (PTH) is a bacterial enzyme that cleaves
peptidyl-tRNA or N-acyl-aminoacyl-tRNA to yield free peptides or N-acyl-amino
acids and tRNA. The natural substrate for this enzyme may be peptidyl-tRNA
which drop off the ribosome during protein synthesis [1,2]. Bacterial PTH has
been found [2,3] to be evolutionary related to yeast hypothetical protein

25 YHR189w.

PTH and YHR189w are proteins of about 200 amino acid residues. As signature
patterns, two conserved regions were selected that each contain an histidine.
The first of these regions is located in the N-terminal section, the other in
the central part.

30 -Consensus pattern: [FY]-x(2)-T-R-H-N-x-G-x(2)-[LIVMFA](2)-[DE]

-Consensus pattern: [GS]-x(3)-H-N-G-[LIVM]-[KR]-[DNS]-[LIVMT]

- [1] Garcia-Villegas M.R., De La Vega F.M., Galindo J.M., Segura M., Buckingham R.H., Guarneros G. EMBO J. 10:3549-3555(1991).
- [2] De La Vega F.M., Galindo J.M., Old I.G., Guarneros G. Gene 169:97-100(1996).
- [3] Ouzounis C., Bork P., Casari G., Sander C. Protein Sci. 4:2424-2428(1995).

5

436. (Peptidase M17) Cytosol aminopeptidase signature

Cytosol aminopeptidase is a eukaryotic cytosolic zinc-dependent exopeptidase that catalyzes the removal of unsubstituted amino-acid residues from the N-terminus of proteins. This enzyme is often known as leucine aminopeptidase (EC 3.4.11.1) (LAP) but has been shown [1] to be identical with prolyl aminopeptidase (EC 3.4.11.5). Cytosol aminopeptidase is a hexamer of identical chains, each of which binds two zinc ions.

Cytosol aminopeptidase is highly similar to Escherichia coli pepA, a manganese dependent aminopeptidase. Residues involved in zinc ion-binding [2] in the mammalian enzyme are absolutely conserved in pepA where they presumably bind manganese.

A cytosol aminopeptidase from Rickettsia prowazekii [3] and one from Arabidopsis thaliana also belong to this family.

As a signature pattern for these enzymes, a perfectly conserved octapeptide was selected which contains two residues involved in binding metal ions: an aspartate and a glutamate.

-Consensus pattern: N-T-D-A-E-G-R-L [The D and the E are zinc/manganese ligands]

-Note: these proteins belong to family M17 in the classification of peptidases [4,E1].

[1] Matsushima M., Takahashi T., Ichinose M., Miki K., Kurokawa K., Takahashi K. Biochem. Biophys. Res. Commun. 178:1459-1464(1991).

[2] Burley S.K., David P.R., Sweet R.M., Taylor A., Lipscomb W.N. J. Mol. Biol. 224:113-140(1992).

[3] Wood D.O., Solomon M.J., Speed R.R. J. Bacteriol. 175:159-165(1993).

[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).

10

15

20

25

30

437. Assemblin (Peptidase family S21)

[1]

Medline: 96399137

5 Three-dimensional structure of human cytomegalovirus
protease.

Shieh HS, Kurumbail RG, Stevens AM, Stegeman RA, Sturman EJ,
Pak JY, Wittwer AJ, Palmier MO, Wiegand RC, Holwerda BC,
Stallings WC;

10 Nature 1996;383:279-282.

Number of members: 29

438. Pollen proteins Ole e I family signature

15 The following plant pollen proteins, whose biological function is not yet
known, are structurally related [1]:

- Olive tree pollen major allergen (Ole e I).
- Tomato anther-specific protein LAT52. - Maize pollen-specific protein ZmC13.

These proteins are most probably secreted and consist of about 145 residues.

20 As shown in the following schematic representation, there are six cysteines
which are conserved in the sequence of these proteins. They seem to be
involved in disulfide bonds.

xxxxxxCxXXXXXXXXXXCXXXXXXXXXXXXXXXXXXXXCXXXXCXXXXXXXXXXXXXXXXXXXXCXXXXXX

*****'C': conserved cysteine involved in a disulfide bond.

25 '*': position of the pattern.

-Consensus pattern: [EQ]-G-x-V-Y-C-D-T-C-R [The two C's are probably involved in
disulfide bonds]

30 [1] Villalba M., Batanero E., Lopez-Otin C., Sanchez L.M., Monsalve R.I., Gonzalez De La
Pena M.A., Lahoz C., Rodriguez R. Eur. J. Biochem. 216:863-869(1993).

439. Pollen allergen

This family contains allergens lol PI, PII and PIII from *Lolium perenne*.

Number of members: 49

[1]

5 Medline: 90105394

Complete primary structure of a *Lolium perenne* (perennial rye grass) pollen allergen, Lol p III: comparison with known Lol p I and II sequences.

Ansari AA, Shenbagamurthi P, Marsh DG;

10 Biochemistry 1989;28:8665-8670.

440. Porphobilinogen deaminase cofactor-binding site

Porphobilinogen deaminase (EC 4.3.1.8), or hydroxymethylbilane synthase, is an enzyme involved in the biosynthesis of porphyrins and related macrocycles. It catalyzes the assembly of four porphobilinogen (PBG) units in a head to tail fashion to form hydroxymethylbilane.

The enzyme covalently binds a dipyrromethane cofactor to which the PBG subunits are added in a stepwise fashion. In the *Escherichia coli* enzyme (gene hemC), this cofactor has been shown [1] to be bound by the sulfur atom of a cysteine. The region around this cysteine is conserved in porphobilinogen deaminases from various prokaryotic and eukaryotic sources.

-Consensus pattern: E-R-x-[LIVMFA]-x(3)-[LIVMF]-x-G-[GSA]-C-x-[IVT]-P-[LIVMF]-
25 [GSA] [C is the cofactor attachment site]

[1] Miller A.D., Hart G.J., Packman L.C., Battersby A.R. Biochem. J. 254:915-918(1988).

30 441. Presenilin

Mutations in presenilin-1 are a major cause of early onset Alzheimer's disease [2]. It has been found that presenilin-1 (Swiss:P49768) binds to beta-catenin in vivo [4]. This family also contains SPE proteins from *C.elegans*.

Number of members: 23

[1]

Medline: 98045995

Presenilins and Alzheimer's disease.

5 Kim TW, Tanzi RE;

Curr Opin Neurobiol 1997;7:683-688.

[2]Medline: 98045995

Presenilins and Alzheimer's disease.

Kim TW, Tanzi RE;

10 Curr Opin Neurobiol 1997;7:683-688.

[3]Medline: 98099802

Interaction of presenilins with the filamin family of
actin-binding proteins.

Zhang W, Han SW, McKeel DW, Goate A, Wu JY;

J Neurosci 1998;18:914-922.

[4]Medline: 99004850

Destabilisation of beta-catenin by mutations in presenilin-1
potentiates neuronal apoptosis.

Zhang Z, Hartmann H, Do VM, Abramowski D, Sturchler-Pierrat

C, Staufenbiel M, Sommer B, van de Wetering M, Clevers H,

Saftig P, De Strooper B, He X, Yankner BA;

Nature 1998;395:698-702.

- 25 442. (Pribosyltran) Purine/pyrimidine phosphoribosyl transferases signature
- Phosphoribosyltransferases (PRT) are enzymes that catalyze the synthesis of
beta-n-5'-monophosphates from phosphoribosylpyrophosphate (PRPP) and an enzyme
specific amine. A number of PRT's are involved in the biosynthesis of purine,
pyrimidine, and pyridine nucleotides, or in the salvage of purines and
30 pyrimidines. These enzymes are:
- Adenine phosphoribosyltransferase (EC 2.4.2.7) (APRT), which is involved in
purine salvage.
 - Hypoxanthine-guanine or hypoxanthine phosphoribosyltransferase (EC 2.4.2.8)

(HGPRT or HPRT), which are involved in purine salvage.

- Orotate phosphoribosyltransferase (EC 2.4.2.10) (OPRT), which is involved in pyrimidine biosynthesis.
- Amido phosphoribosyltransferase (EC 2.4.2.14), which is involved in purine biosynthesis.
- Xanthine-guanine phosphoribosyltransferase (EC 2.4.2.22) (XGPRT), which is involved in purine salvage.

In the sequence of all these enzymes there is a small conserved region which may be involved in the enzymatic activity and/or be part of the PRPP binding site [1].

-Consensus pattern: [LIVMFYWCTA]-[LIVM]-[LIVMA]-[LIVMFC]-[DE]-D-[LIVMS]-[LIVM]-[STAVD]-[STAR]-[GAC]-x-[STAR]

-Note: in position 11 of the pattern most of these enzymes have Gly.

[1] Hershey H.V., Taylor M.W. Gene 43:287-293(1986).

443. (Pro CA)

Prokaryotic-type carbonic anhydrases signatures

Carbonic anhydrases (EC 4.2.1.1) (CA) are zinc metalloenzymes which catalyze the reversible hydration of carbon dioxide. In *Escherichia coli*, CA (gene *cynT*) is involved in recycling carbon dioxide formed in the bicarbonate-dependent decomposition of cyanate by cyanase (gene *cynS*). By this action, it prevents the depletion of cellular bicarbonate [1]. In photosynthetic bacteria and plant chloroplast, CA is essential to inorganic carbon fixation [2]. Prokaryotic and plant chloroplast CA are structurally and evolutionary related and form a family distinct from the one which groups the many different forms of eukaryotic CA's (see <PDOC00146>). Hypothetical proteins *yadF* from *Escherichia coli* and HI1301 from *Haemophilus influenzae* also belong to this family. Two signature patterns were developed for this family of enzymes. Both patterns contain conserved residues that could be involved in binding zinc (cysteine and histidine).

-Consensus pattern: C-[SA]-D-S-R-[LIVM]-x-[AP]

-Consensus pattern: [EQ]-Y-A-[LIVM]-x(2)-[LIVM]-x(4)-[LIVMF](3)-x-G-H-x(2)-C-G

[1] Guilloton M.B., Korte J.J., Lamblin A.F., Fuchs J.A., Anderson P.M. J. Biol. Chem. 267:3731-3734(1992).

[2] Fukuzawa H., Suzuki E., Komukai Y., Miyachi S. Proc. Natl. Acad. Sci. U.S.A. 89:4437-4441(1992).

444. (Prolyl_oligopep)

Prolyl oligopeptidase family serine active site

The prolyl oligopeptidase family [1,2,3] consist of a number of evolutionary related peptidases whose catalytic activity seems to be provided by a charge relay system similar to that of the trypsin family of serine proteases, but which evolved by independent convergent evolution. The known members of this family are listed below.

- Prolyl endopeptidase (EC 3.4.21.26) (PE) (also called post-proline cleaving enzyme). PE is an enzyme that cleaves peptide bonds on the C-terminal side of prolyl residues. The sequence of PE has been obtained from a mammalian species (pig) and from bacteria (*Flavobacterium meningosepticum* and *Aeromonas hydrophila*); there is a high degree of sequence conservation between these sequences.

- *Escherichia coli* protease II (EC 3.4.21.83) (oligopeptidase B) (gene prtB) which cleaves peptide bonds on the C-terminal side of lysyl and arginyl residues.

- Dipeptidyl peptidase IV (EC 3.4.14.5) (DPP IV). DPP IV is an enzyme that removes N-terminal dipeptides sequentially from polypeptides having unsubstituted N-termini provided that the penultimate residue is proline.

- Yeast vacuolar dipeptidyl aminopeptidase A (DPAP A) (gene: STE13) which is responsible for the proteolytic maturation of the alpha-factor precursor.

- Yeast vacuolar dipeptidyl aminopeptidase B (DPAP B) (gene: DAP2).

- Acylamino-acid-releasing enzyme (EC 3.4.19.1) (acyl-peptide hydrolase).

This enzyme catalyzes the hydrolysis of the amino-terminal peptide bond of an N-acetylated protein to generate a N-acetylated amino acid and a protein with a free amino-terminus.

A conserved serine residue has experimentally been shown (in E.coli proteaseII as well as in pig and bacterial PE) to be necessary for the catalytic mechanism. This serine, which is part of the catalytic triad (Ser, His, Asp), is generally located about 150 residues away from the C-terminal extremity of these enzymes (which are all proteins that contains about 700 to 800 amino acids).

Consensus pattern: D-x(3)-A-x(3)-[LIVMFYW]-x(14)-G-x-S-x-G-G-[LIVMFYW](2) [S is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for yeast DPAP A.

Note: these proteins belong to families S9A/S9B/S9C in the classification of peptidases [4].

[1] Rawlings N.D., Polgar L., Barrett A.J. Biochem. J. 279:907-911(1991).

□

[2] Barrett A.J., Rawlings N.D.

□

[3] Polgar L., Szabo E.

□

[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

445. (Pterin 4a)

Pterin 4 alpha carbinolamine dehydratase

Pterin 4 alpha carbinolamine dehydratase is aka DCoH (dimerisation cofactor of hepatocyte nuclear factor 1-alpha).

Number of members: 11

[1] Cronk JD, Endrizzi JA, Alber T; Medline: 97052967 "High-resolution structures of the bifunctional enzyme and transcriptional coactivator DCoH and its complex with a product analogue." Protein Sci 1996;5:1963-1972.

446. (Pyridox oxidase)

Pyridoxamine 5'-phosphate oxidase signature

5

Pyridoxamine 5'-phosphate oxidase (EC 1.4.3.5) is a FMN flavoprotein involved in the de novo synthesis of pyridoxine (vitamin B6) and pyridoxal phosphate. It oxidizes pyridoxamine-5-P (PMP) and pyridoxine-5-P (PNP) to pyridoxal-5-P. The sequences of the enzyme from bacterial (genes *pdxH* or *fprA*) [1] and fungal (gene *PDX3*) [2] sources show that this protein has been highly conserved throughout evolution.

10

PdxH is evolutionary related [3] to one of the enzymes in the phenazine biosynthesis protein pathway, *phzD* (also known as *phzG*). As a signature pattern, a highly conserved region was selected located in the C-terminal part of these enzymes.

-Consensus pattern: [LIVF]-E-F-W-[QHG]-x(4)-R-[LIVM]-H-[DNE]-R

[1] Lam H.-M., Winkler M.E. J. Bacteriol. 174:6033-6045(1992).

[2] Loubbardi A., Karst F., Guilloton M., Marcireau C. J. Bacteriol. 177:1817-1823(1995).

[3] Pierson L.S. III, Gaffney T., Lam S., Gong F. FEMS Microbiol. Lett. 134:299-307(1995).

20

447. (Pyrophosphatase)

Inorganic pyrophosphatase signature

25

Inorganic pyrophosphatase (EC 3.6.1.1) (PPase) [1,2] is the enzyme responsible for the hydrolysis of pyrophosphate (PPi) which is formed principally as the product of the many biosynthetic reactions that utilize ATP. All known Ppases require the presence of divalent metal cations, with magnesium conferring the highest activity. Among other residues, a lysine has been postulated to be part or close to the active site. PPases have been sequenced from bacteria such as *Escherichia coli* (homohexamer), thermophilic bacteria PS-3 and *Thermus thermophilus*, from the archaebacteria *Thermoplasma acidophilum*, from fungi (homodimer), from a plant, and from bovine retina. In yeast, a mitochondrial isoform of

30

PPase has been characterized which seems to be involved in energy production and whose activity is stimulated by uncouplers of ATP synthesis.

The sequences of PPases share some regions of similarities. As signature patterns a region was selected that contains three conserved aspartates that are involved in the binding of cations.

-Consensus pattern: D-[SGDN]-D-[PE]-[LIVMF]-D-[LIVMGAC]
[The three D's bind divalent metal cations]

[1] Lahti R., Kolakowski L.F. Jr., Heinonen J., Vihinen M., Pohjanoksa K., Cooperman B.S. Biochim. Biophys. Acta 1038:338-345(1990).

[2] Cooperman B.S., Baykov A.A., Lahti R. Trends Biochem. Sci. 17:262-266(1992).

448. (Peptidase S26)

Signal peptidases I signatures.

Signal peptidases (SPases) [1] (aka leader peptidases) remove the signal peptides from secretory proteins. In prokaryotes three types of SPases are known: type I (gene *lepB*) which is responsible for the processing of the majority of exported pre-proteins; type II (gene *lsp*) which only process lipoproteins, and a third type involved in the processing of pili subunits. SPase I (EC 3.4.21.89) is an integral membrane protein that is anchored in the cytoplasmic membrane by one (in *B. subtilis*) or two (in *E. coli*) N-terminal transmembrane domains with the main part of the protein protruding in the periplasmic space. Two residues have been shown [2,3] to be essential for the catalytic activity of SPase I: a serine and an lysine. SPase I is evolutionary related to the yeast mitochondrial inner membrane protease subunit 1 and 2 (genes *IMP1* and *IMP2*) which catalyze the removal of signal peptides required for the targeting of proteins from the mitochondrial matrix, across the inner membrane, into the inter-membrane space [4]. In eukaryotes the removal of signal peptides is effected by an oligomeric enzymatic complex composed of at least five subunits: the signal peptidase complex (SPC). The SPC is located in the endoplasmic reticulum membrane. Two components of mammalian SPC, the 18 Kd (SPC18) and the 21 Kd (SPC21) subunits as well

as the yeast SEC11 subunit have been shown [5] to share regions of sequence similarity with prokaryotic SPases I and yeast IMP1/IMP2. Three signature patterns have been developed for these proteins. The first signature contains the putative active site serine, the second signature contains the putative active site lysine which is not conserved in the SPC subunits, and the third signature corresponds to a conserved region of unknown biological significance which is located in the C-terminal section of all these proteins.

Consensus pattern: [GS]-x-S-M-x-[PS]-[AT]-[LF] [S is an active site residue]-

Consensus pattern: K-R-[LIVMSTA](2)-G-x-[PG]-G-[DE]-x-[LIVM]-x-[LIVMFY] [K is an active site residue]-

Consensus pattern: [LIVMFYW](2)-x(2)-G-D-[NH]-x(3)-[SND]-x(2)-[SG]-

[1] Dalbey R.E., von Heijne G. Trends Biochem. Sci. 17:474-478(1992).[2] Sung M., Dalbey R.E. J. Biol. Chem. 267:13154-13159(1992).[3] Black M.T. J. Bacteriol. 175:4957-4961(1993).[4] Nunnari J., Fox T.D., Walter P. Science 262:1997-2004(1993).[5] van Dijk J.M., de Jong A., Vehmaanpera J., Venema G., Bron S. EMBO J. 11:2819-2828(1992).[6] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).[E1]

449. (Peptidase C1) Eukaryotic thiol (cysteine) proteases active sites. Eukaryotic thiol proteases (EC 3.4.22.-) [1] are a family of proteolytic enzymes which contain an active site cysteine. Catalysis proceeds through a thioester intermediate and is facilitated by a nearby histidine side chain; an asparagine completes the essential catalytic triad. The proteases which are currently known to belong to this family are listed below (references are only provided for recently determined sequences). - Vertebrate lysosomal cathepsins B (EC 3.4.22.1), H (EC 3.4.22.16), L (EC 3.4.22.15), and S (EC 3.4.22.27) [2]. - Vertebrate lysosomal dipeptidyl peptidase I (EC 3.4.14.1) (also known as cathepsin C) [2]. - Vertebrate calpains (EC 3.4.22.17). Calpains are intracellular calcium- activated thiol protease that contain both a N-terminal catalytic domain and a C-terminal calcium-binding domain. - Mammalian cathepsin K, which seems involved in osteoclastic bone resorption [3]. - Human cathepsin O [4]. - Bleomycin hydrolase. An enzyme that catalyzes the inactivation of the antitumor drug BLM (a glycopeptide). - Plant enzymes: barley aleurain (EC 3.4.22.16), EP-B1/B4; kidney bean EP-C1, rice bean SH-EP; kiwi fruit actinidin (EC 3.4.22.14); papaya

latex papain (EC 3.4.22.2), chymopapain (EC 3.4.22.6), caricain (EC 3.4.22.30), and proteinase IV (EC 3.4.22.25); pea turgor-responsive protein 15A; pineapple stem bromelain (EC 3.4.22.32); rape COT44; rice oryzain alpha, beta, and gamma; tomato low-temperature induced, *Arabidopsis thaliana* A494, RD19A and RD21A. - House-dust mites allergens DerP1 and EurM1. - Cathepsin B-like proteinases from the worms *Caenorhabditis elegans* (genes *gcp-1*, *cpr-3*, *cpr-4*, *cpr-5* and *cpr-6*), *Schistosoma mansoni* (antigen SM31) and *Japonica* (antigen SJ31), *Haemonchus contortus* (genes AC-1 and AC-2), and *Ostertagia ostertagi* (CP-1 and CP-3). - Slime mold cysteine proteinases CP1 and CP2. - Cruzipain from *Trypanosoma cruzi* and *brucei*. - Throphozoite cysteine proteinase (TCP) from various *Plasmodium* species. - Proteases from *Leishmania mexicana*, *Theileria annulata* and *Theileria parva*. - Baculoviruses cathepsin-like enzyme (v-cath). - *Drosophila* small optic lobes protein (gene *sol*), a neuronal protein that contains a calpain-like domain. - Yeast thiol protease BLH1/YCP1/LAP3. - *Caenorhabditis elegans* hypothetical protein C06G4.2, a calpain-like protein. Two bacterial peptidases are also part of this family: - Aminopeptidase C from *Lactococcus lactis* (gene *pepC*) [5]. - Thiol protease *tpr* from *Porphyromonas gingivalis*. Three other proteins are structurally related to this family, but may have lost their proteolytic activity. - Soybean oil body protein P34. This protein has its active site cysteine replaced by a glycine. - Rat testin, a sertoli cell secretory protein highly similar to cathepsin L but with the active site cysteine is replaced by a serine. Rat testin should not be confused with mouse testin which is a LIM-domain protein (see <[PDOC00382](#)>). - *Plasmodium falciparum* serine-repeat protein (SERA), the major blood stage antigen. This protein of 111 Kd possesses a C-terminal thiol-protease-like domain [6], but the active site cysteine is replaced by a serine. The sequences around the three active site residues are well conserved and can be used as signature patterns.

Consensus pattern: Q-x(3)-[GE]-x-C-[YW]-x(2)-[STAGC]-[STAGCV] [C is the active site residue]- Note: the residue in position 4 of the pattern is almost always cysteine; the only exceptions are calpains (Leu), bleomycin hydrolase (Ser) and yeast YCP1 (Ser). -Note: the residue in position 5 of the pattern is always Gly except in papaya protease IV where it is Glu.

Consensus pattern: [LIVMGSTAN]-x-H-[GSACE]-[LIVM]-x-[LIVMAT](2)-G-x-[GSADNH] [H is the active site residue]-

Consensus pattern: [FYCH]-[WI]-[LIVT]-x-[KRQAG]-N-[ST]-W-x(3)-[FYW]-G-x(2)-G-[LFYW]-[LIVMFYG]-x-[LIVMF] [N is the active site residue] - Note: these proteins belong to family C1 (papain-type) and C2 (calpains) in the classification of peptidases [7,E1].-

- 5 [1] Dufour E. Biochimie 70:1335-1342(1988).[2] Kirschke H., Barrett A.J., Rawlings N.D. Protein Prof. 2:1587-1643(1995).[3] Shi G.-P., Chapman H.A., Bhairi S.M., Deleeuw C., Reddy V.Y., Weiss S.J. FEBS Lett. 357:129-134(1995).[4] Velasco G., Ferrando A.A., Puente X.S., Sanchez L.M., Lopez-Otin C. J. Biol. Chem. 269:27136-27142(1994).[5] Chapot-Chartier M.P., Nardi M., Chopin M.C., Chopin A., Gripon J.C. Appl. Environ. Microbiol. 59:330-333(1993).[6] Higgins D.G., McConnell D.J., Sharp P.M. Nature 340:604-604(1989).[7] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:461-486(1994).

450. (peptidase M24) Aminopeptidase P and proline dipeptidase signature (1).

- 15 Aminopeptidase P (EC 3.4.11.9) is the enzyme responsible for the release of any N-terminal amino acid adjacent to a proline residue. Proline dipeptidase(EC 3.4.13.9) (prolidase) splits dipeptides with a prolyl residue in the carboxyl terminal position. Bacterial aminopeptidase P II (gene pepP) [1], proline dipeptidase (gene pepQ)[2], and human proline dipeptidase (gene PEPD) [3] are evolutionary related. These proteins are manganese metalloenzymes. Yeast 20 hypothetical proteins YER078c and YFR006w and Mycobacterium tuberculosis hypothetical protein MtCY49.29c also belong to this family. As a signature pattern for these enzymes a conserved region that contains three histidine residues has been developed

Consensus pattern: [HA]-[GSYR]-[LIVMT]-[SG]-H-x-[LIV]-G-[LIVM]-x-[IV]-H-[DE]-

- 25 [1] Yoshimoto T., Tone H., Honda T., Osatomi K., Kobayashi R., Tsuru D. J. Biochem. 105:412-416(1989).[2] Nakahigashi K., Inokuchi H. Nucleic Acids Res. 18:6439-6439(1990).[3] Endo F., Tanoue A., Nakai H., Hata A., Indo Y., Titani K., Matsuda I. J. Biol. Chem. 264:4476-4481(1989).[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-30 228(1995).

Methionine aminopeptidase signatures. (2). Methionine aminopeptidase (EC 3.4.11.18) (MAP) is responsible for the removal of the amino-terminal (initiator) methionine from

nascent eukaryotic cytosolic and cytoplasmic prokaryotic proteins if the penultimate amino acid is small and uncharged. All MAP studied to date are monomeric proteins that require cobalt ions for activity. Two subfamilies of MAP enzymes are known to exist [1,2]. While being evolutionary related, they only share a limited amount of sequence similarity mostly clustered around the residues shown, in the Escherichia coli MAP [3], to be involved in cobalt-binding. The first family consists of enzymes from prokaryotes as well as eukaryotic MAP-1, while the second group is made up of archebacterial MAP and eukaryotic MAP-2. The second subfamily also includes proteins which do not seem to be MAP, but that are clearly evolutionary related such as mouse proliferation-associated protein 1 and fission yeast curved DNA-binding protein. For each of these subfamilies, a specific signature pattern that includes residues known to be involved in cobalt-binding has been developed.

Consensus pattern: [MFY]-x-G-H-G-[LIVMC]-[GSH]-x(3)-H-x(4)-[LIVM]-x-[HN]-[YWV]
[H is a cobalt ligand]-

Consensus pattern: [DA]-[LIVMY]-x-K-[LIVM]-D-x-G-x-[HQ]-[LIVM]-[DNS]-G-x(3)-
[DN] [The second D and the last D/N are cobalt ligands]

[1] Arfin S.M., Kendall R.L., Hall L., Weaver L.H., Stewart A.E., Matthews B.W.,
Bradshaw R.A. Proc. Natl. Acad. Sci. U.S.A. 92:7714-7718(1995).[2] Keeling P.J., Doolittle
W.F. Trends Biochem. Sci. 21:285-286(1996).[3] Roderick S.L., Mathews B.W.
Biochemistry 32:3907-3912(1993).[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-
228(1995).

451. Cytochrome P450 cysteine heme-iron ligand signature

Cytochrome P450's [1,2,3,E1] are a group of enzymes involved in the oxidative metabolism of a high number of natural compounds (such as steroids, fatty acids, prostaglandins, leukotrienes, etc) as well as drugs, carcinogens and mutagens. Based on sequence similarities, P450's have been classified into about forty different families [4,5]. P450's are proteins of 400 to 530 amino acids; the only exception is Bacillus BM-3 (CYP102) which is a protein of 1048 residues that contains a N-terminal P450 domain followed by a reductase domain. P450's are heme proteins. A conserved cysteine residue in the C-terminal part of P450's is

involved in binding the heme iron in the fifth coordination site. From a region around this residue, a ten residue signature was developed specific to P450's.

Consensus pattern: [FW]-[SGNH]-x-[GD]-x-[RHPT]-x-C-[LIVMFAP]-[GAD] [C is the heme iron ligand]-

[1] Nebert D.W., Gonzalez F.J. Annu. Rev. Biochem. 56:945-993(1987).

[2] Coon M.J., Ding X., Pernecky S.J., Vaz A.D.N. FASEB J. 6:669-673(1992).

[3] Guengerich F.P. J. Biol. Chem. 266:10019-10022(1991).

[4] Nelson D.R., Kamataki T., Waxman D.J., Guengerich F.P., Estrabrook R.W., Feyereisen R., Gonzalez F.J., Coon M.J., Gunsalus I.C., Gotoh O., Okuda K., Nebert D.W. DNA Cell Biol. 12:1-51(1993).

[5] Degtyarenko K.N., Archakov A.I. FEBS Lett. 332:1-8(1993).

452. (Pec Lyase) Pectate lyase

This enzyme forms a right handed beta helix structure. Pectate lyase is an enzyme involved in the maceration and soft rotting of plant tissue.

[1] Yoder MD, Keen NT, Jurnak F, Science 1993;260:1503-1507.

453. (pep M24) Aminopeptidase P and proline dipeptidase signature (pep1)

Aminopeptidase P (EC 3.4.11.9) is the enzyme responsible for the release of any N-terminal amino acid adjacent to a proline residue. Proline dipeptidase(EC 3.4.13.9) (prolidase) splits dipeptides with a prolyl residue in the carboxyl terminal position. Bacterial aminopeptidase P II (gene pepP) [1], proline dipeptidase (gene pepQ)[2], and human proline dipeptidase (gene PEPP) [3] are evolutionary related. These proteins are manganese metalloenzymes. Yeast hypothetical proteins YER078c and YFR006w and Mycobacterium tuberculosis .hypothetical protein MtCY49.29c also belong to this family. As a signature pattern for these enzymes a conserved region was selected that contains three histidine residues.

Consensus pattern: [HA]-[GSYR]-[LIVMT]-[SG]-H-x-[LIV]-G-[LIVM]-x-[IV]-H-[DE]-

- [1] Yoshimoto T., Tone H., Honda T., Osatomi K., Kobayashi R., Tsuru D. J. Biochem. 105:412-416(1989).
- [2] Nakahigashi K., Inokuchi H. Nucleic Acids Res. 18:6439-6439(1990).
- [3] Endo F., Tanoue A., Nakai H., Hata A., Indo Y., Titani K., Matsuda I. J. Biol. Chem. 264:4476-4481(1989).
- [4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).

Methionine aminopeptidase signatures (pep2)

Methionine aminopeptidase (EC 3.4.11.18) (MAP) is responsible for the removal of the amino-terminal (initiator) methionine from nascent eukaryotic cytosolic and cytoplasmic prokaryotic proteins if the penultimate amino acid is small and uncharged. All MAP studied to date are monomeric proteins that require cobalt ions for activity. Two subfamilies of MAP enzymes are known to exist [1,2]. While being evolutionary related, they only share a limited amount of sequence similarity mostly clustered around the residues shown, in the Escherichia coli MAP [3], to be involved in cobalt-binding. The first family consists of enzymes from prokaryotes as well as eukaryotic MAP-1, while the second group is made up of archeobacterial MAP and eukaryotic MAP-2. The second subfamily also includes proteins which do not seem to be MAP, but that are clearly evolutionary related such as mouse proliferation-associated protein 1 and fission yeast curved DNA-binding protein. For each of these subfamilies, a specific signature pattern was developed that includes residues known to be involved in cobalt-binding.

Consensus pattern: [MFY]-x-G-H-G-[LIVMC]-[GSH]-x(3)-H-x(4)-[LIVM]-x-[HN]- [YWV]
[H is a cobalt ligand]-

Consensus pattern: [DA]-[LIVMY]-x-K-[LIVM]-D-x-G-x-[HQ]-[LIVM]-[DNS]-G-x(3)-[DN] [The second D and the last D/N are cobalt ligands]

- [1] Arfin S.M., Kendall R.L., Hall L., Weaver L.H., Stewart A.E., Matthews B.W., Bradshaw R.A. Proc. Natl. Acad. Sci. U.S.A. 92:7714-7718(1995).
- [2] Keeling P.J., Doolittle W.F. Trends Biochem. Sci. 21:285-286(1996).
- [3] Roderick S.L., Mathews B.W. Biochemistry 32:3907-3912(1993).
- [4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).

454. Peroxidases signatures

Peroxidases (EC 1.11.1.-) [1] are heme-binding enzymes that carry out a variety of biosynthetic and degradative functions using hydrogen peroxide as the electron acceptor.

5 Peroxidases are widely distributed throughout bacteria, fungi, plants, and vertebrates. In peroxidases the heme prosthetic group is protoporphyrin IX and the fifth ligand of the heme iron is a histidine (known as the proximal histidine). Another histidine residue (the distal histidine) serves as an acid-base catalyst in the reaction between hydrogen peroxide and the enzyme. The regions around these two active site residues are more or less conserved in a majority of peroxidases [2,3]. The enzymes in which one or both of these regions can be found are listed below. - Yeast cytochrome c peroxidase (EC 1.11.1.5). - Myeloperoxidase (EC 1.11.1.7) (MPO). MPO is found in granulocytes and monocytes and plays a major role in the oxygen-dependent microbicidal system of neutrophils. - Lactoperoxidase (EC 1.11.1.7) (LPO). LPO is a milk protein which acts as an antimicrobial agent. - Eosinophil peroxidase (EC 1.11.1.7) (EPO). An enzyme found in the cytoplasmic granules of eosinophils. - Thyroid peroxidase (EC 1.11.1.8) (TPO). TPO plays a central role in the biosynthesis of thyroid hormones. It catalyzes the iodination and coupling of the hormonogenic tyrosines in thyroglobulin to yield the thyroid hormones T3 and T4. - Fungal ligninases. Ligninase catalyzes the first step in the degradation of lignin. It depolymerizes lignin by catalyzing the C(alpha)-C(beta) cleavage of the propyl side chains of lignin. - Plant peroxidases (EC 1.11.1.7). Plants express a large number of isozymes of peroxidases. Some of them play a role in cell-suberization by catalyzing the deposition of the aromatic residues of suberin on the cell wall, some are expressed as a defense response toward wounding, others are involved in the metabolism of auxin and the biosynthesis of lignin. - Prokaryotic catalase-peroxidases. 25 Some bacterial species produce enzymes that exhibit both catalase and broad-spectrum peroxidase activities [4]. Examples of such enzymes are: catalase HP I from *Escherichia coli* (gene *katG*) and *perA* from *Bacillus stearothermophilus*.

Consensus pattern: [DET]-[LIVMTA]-x(2)-[LIVM]-[LIVMSTAG]-[SAG]-[LIVMSTAG]-H-[STA]-[LIVMFY] [H is the proximal heme-binding ligand] -

Consensus pattern: [SGATV]-x(3)-[LIVMA]-R-[LIVMA]-x-[FW]-H-x-[SAC] [H is an active site residue]-

- [1] Dawson J.H. Science 240:433-439(1988).
- [2] Kimura S., Ikeda-Saito M. Proteins 3:113-120(1988).
- [3] Henrissat B., Saloheimo M., Lavaitte S., Knowles J.K.C. Proteins 8:251-257(1990).
- [4] Welinder K.G. Biochim. Biophys. Acta 1080:215-220(1991).

5

455. pfkB family of carbohydrate kinases signatures

It has been shown [1,2,3] that the following carbohydrate and purine kinases are evolutionary related and can be grouped into a single family, which is known [1] as the 'pfkB family': -

- 10 Fructokinase (EC 2.7.1.4) (gene scrK). - 6-phosphofructokinase isozyme 2 (EC 2.7.1.11) (phosphofructokinase-2) (gene pfkB). pfkB is a minor phosphofructokinase isozyme in Escherichia coli and is not evolutionary related to the major isozyme (gene pfkA). Plants 6-phosphofructokinase also belong to this family. - Ribokinase (EC 2.7.1.15) (gene rbsK). - Adenosine kinase (EC 2.7.1.20) (gene ADK). - 2-dehydro-3-deoxygluconokinase (EC 2.7.1.45) (gene: kdgK). - 1-phosphofructokinase (EC 2.7.1.56) (fructose 1-phosphate kinase) (gene fruK). - Inosine-guanosine kinase (EC 2.7.1.73) (gene gsk). - Tagatose-6-phosphate kinase (EC 2.7.1.144) (phosphotagatokinase) (gene lacC). - Escherichia coli hypothetical protein yeiC. - Escherichia coli hypothetical protein yeiI. - Escherichia coli hypothetical protein yhfQ. - Escherichia coli hypothetical protein yihV. - Bacillus subtilis hypothetical protein yxdC. - Yeast hypothetical protein YJR105w. All the above kinases are proteins of from 280 to 430 amino acid residues that share a few region of sequence similarity. Two of these regions were selected as signature patterns. The first pattern is based on a region rich in glycine which is located in the N-terminal section of these enzymes; while the second pattern is based on a conserved region in the C-terminal section.

25

Consensus pattern: [AG]-G-x(0,1)-[GAP]-x-N-x-[STA]-x(6)-[GS]-x(9)-G-

Consensus pattern: [DNSK]-[PSTV]-x-[SAG](2)-[GD]-D-x(3)-[SAGV]-[AG]-[LIVMFYA]-[LIVMSTAP]

- 30 [1] Wu L.-F., Reizer A., Reizer J., Cai B., Tomich J.M., Saier M.H. Jr. J. Bacteriol. 173:3117-3127(1991).
- [2] Orchard L.M.D., Kornberg H.L. Proc. R. Soc. Lond., B, Biol. Sci. 242:87-90(1990).
- [3] Blatch G.L., Scholle R.R., Woods D.R. Gene 95:17-23(1990).

456. Phospholipase A2 active sites signatures

Phospholipase A2 (EC 3.1.1.4) (PA2) [1,2] is an enzyme which releases fatty acids from the second carbon group of glycerol. PA2's are small and rigid proteins of 120 amino-acid residues that have four to seven disulfide bonds. PA2 binds a calcium ion which is required for activity. The side chains of two conserved residues, a histidine and an aspartic acid, participate in a 'catalytic network'. Many PA2's have been sequenced from snakes, lizards, bees and mammals. In the latter, there are at least four forms: pancreatic, membrane-associated as well as two less characterized forms. The venom of most snakes contains multiple forms of PA2. Some of them are presynaptic neurotoxins which inhibit neuromuscular transmission by blocking acetylcholine release from the nerve termini. Two different signature patterns were derived for PA2's. The first is centered on the active site histidine and contains three cysteines involved in disulfide bonds. The second is centered on the active site aspartic acid and also contains three cysteines involved in disulfide bonds.

Consensus pattern: C-C-x(2)-H-x(2)-C [H is the active site residue] This pattern will not detect some snake toxins homologous with PA2 but which have lost their catalytic activity as well as otoconin-22, a *Xenopus* protein from the aragonitic otoconia which is also unlikely to be enzymatically active.

Consensus pattern: [LIVMA]-C-{LIVMFYWPCST}-C-D-x(5)-C [D is the active site residue] The majority of functional and non-functional PA2's. Undetected sequences are bee PA2, gila monster PA2's, PA2 PL-X from habu and PA2 PA-5 from mulga.

[1] Davidson F.F., Dennis E.A. J. Mol. Evol. 31:228-238(1990).

[2] Gomez F., Vandermeers A., Vandermeers-Piret M.-C., Herzog R., Rathe J., Stievenart M., Winand J., Christophe J. Eur. J. Biochem. 186:23-33(1989).

457. Phosphorylase pyridoxal-phosphate attachment site. Phosphorylases (EC 2.4.1.1) [1] are important allosteric enzymes in carbohydrate metabolism. They catalyze the formation of glucose 1-phosphate from polyglucose such as glycogen, starch or maltodextrin. Enzymes from different sources differ in their regulatory mechanisms and their natural substrates.

However, all known phosphorylases share catalytic and structural properties. They are pyridoxal-phosphate dependent enzymes; the pyridoxal-P group is attached to a lysine residue around which the sequence is highly conserved and can be used as a signature pattern to detect this class of enzymes.

5

Consensus pattern: E-A-[SC]-G-x-[GS]-x-M-K-x(2)-[LM]-N [K is the pyridoxal-P attachment site]-

[1] Fukui T., Shimomura S., Nakano K. Mol. Cell. Biochem. 42:129-144(1982).

10

458. Protein kinases signatures and profile

Eukaryotic protein kinases [1 to 5] are enzymes that belong to a very extensive family of proteins which share a conserved catalytic core common to both serine/threonine and tyrosine protein kinases. There are a number of conserved regions in the catalytic domain of protein kinases. Two of these regions were selected to build signature patterns. The first region, which is located in the N-terminal extremity of the catalytic domain, is a glycine-rich stretch of residues in the vicinity of a lysine residue, which has been shown to be involved in ATP binding. The second region, which is located in the central part of the catalytic domain, contains a conserved aspartic acid residue which is important for the catalytic activity of the enzyme [6]; Two signature patterns were derived for that region: one specific for serine/threonine kinases and the other for tyrosine kinases. A profile was also developed which is based on the alignment in [1] and covers the entire catalytic domain.

25 Consensus pattern: [LIV]-G-{P}-G-{P}-[FYWMGSTNH]-[SGA]-{PW}-[LIVCAT]-{PD}-x-[GSTACLIVMFY]-x(5,18)-[LIVMFYWCSTAR]-[AIVP]-[LIVMFAGCKR]-K [K binds ATP]. The majority of known protein kinases belong to the class detected by this pattern, but it fails to find a number of them, especially viral kinases which are quite divergent in this region and are completely missed by this pattern.

30 Consensus pattern: [LIVMFYC]-x-[HY]-x-D-[LIVMFY]-K-x(2)-N-[LIVMFYCT](3) [D is an active site residue]. Most serine/ threonine specific protein kinases belong to this class detected by the pattern with 10 exceptions (half of them viral kinases) and also Epstein-Barr virus BGLF4 and Drosophila ninaC which have respectively Ser and Arg instead of the

conserved Lys and which are therefore detected by the tyrosine kinase specific pattern described below.

Consensus pattern: [LIVMFYC]-x-[HY]-x-D-[LIVMFY]-[RSTAC]-x(2)-N-[LIVMFYC](3)

[D is an active site residue] ALL tyrosine specific protein kinases with the exception of human ERBB3 and mouse blk belong to this class detected by the pattern. This pattern will also detect most bacterial aminoglycoside phosphotransferases [8,9] and herpesviruses gangciclovir kinases [10]; which are proteins structurally and evolutionary related to protein kinases. This profile also detects receptor guanylate cyclases and 2-5A-dependent ribonucleases. Sequence similarities between these two families and the eukaryotic protein kinase family have been noticed before. It also detects Arabidopsis thaliana kinase- like protein TMKL1 which seems to have lost its catalytic activity. If a protein analyzed includes the two protein kinase signatures, the probability of it being a protein kinase is close to 100%. Eukaryotic-type protein kinases have also been found in prokaryotes such as Myxococcus xanthus [11] and Yersinia pseudotuberculosis.

[1] Hanks S.K., Hunter T. FASEB J. 9:576-596(1995).

[2] Hunter T. Meth. Enzymol. 200:3-37(1991).

[3] Hanks S.K., Quinn A.M. Meth. Enzymol. 200:38-62(1991).

[4] Hanks S.K. Curr. Opin. Struct. Biol. 1:369-383(1991).

[5] Hanks S.K., Quinn A.M., Hunter T. Science 241:42-52(1988).

[6] Knighton D.R., Zheng J., Ten Eyck L.F., Ashford V.A., Xuong N.-H., Taylor S.S., Sowadski J.M. Science 253:407-414(1991).

[7] Bairoch A., Claverie J.-M. Nature 331:22(1988).

[8] Benner S. Nature 329:21-21(1987).

[9] Kirby R. J. Mol. Evol. 30:489-492(1992).

[10] Littler E., Stuart A.D., Chee M.S. Nature 358:160-162(1992).

[11] Munoz-Dorado J., Inouye S., Inouye M. Cell 67:995-1006(1991).

Receptor tyrosine kinase class II signature

A number of growth factors stimulate mitogenesis by interacting with a family of cell surface receptors which possess an intrinsic, ligand-sensitive, protein tyrosine kinase activity [1]. These receptor tyrosine kinases (RTK) all share the same topology: an extracellular ligand-binding domain, a single transmembrane region and a cytoplasmic kinase domain. However

they can be classified into at least five groups. The prototype for class II RTK's is the insulin receptor, a heterotetramer of two alpha and two beta chains linked by disulfide bonds. The alpha and beta chains are cleavage products of a precursor molecule. The alpha chain contains the ligand binding site, the beta chain transverses the membrane and contains the tyrosine protein kinase domain. The receptors currently known to belong to class II are: - Insulin receptor from vertebrates. - Insulin growth factor I receptor from mammals. - Insulin receptor-related receptor (IRR), which is most probably a receptor for a peptide belonging to the insulin family. - Insects insulin-like receptors. - Molluscan insulin-related peptide(s) receptor (MIP-R). - Insulin-like peptide receptor from Branchiostoma lanceolatum. - The Drosophila developmental protein sevenless, a putative receptor for positional information required for the formation of the R7 photoreceptor cells. - The trk family of receptors (NTRK1, NTRK2 and NTRK3), which are high affinity receptors for nerve growth factor and related neurotrophic factors (BDNF and NT-3). And the following uncharacterized receptors: - ROS. - LTK (TYK1). - EDDR1 (cak, TRKE, RTK6). - NTRK3 (Tyro10, TKT). - A sponge putative receptor tyrosine kinase. While only the insulin and the insulin growth factor I receptors are known to exist in the tetrameric conformation specific to class II RTK's, all the above proteins share extensive homologies in their kinase domain, especially around the putative site of autophosphorylation. Hence, a signature pattern was developed for this class of RTK's, which includes the tyrosine residue, itself probably autophosphorylated.

Consensus pattern: [DN]-[LIV]-Y-x(3)-Y-Y-R [The second Y is the autophosphorylation site]

[1] Yarden Y., Ullrich A. Annu. Rev. Biochem. 57:443-478(1988).

Receptor tyrosine kinase class III signature

A number of growth factors stimulate mitogenesis by interacting with a family of cell surface receptors which possess an intrinsic, ligand-sensitive, protein tyrosine kinase activity [1].

These receptor tyrosine kinases (RTK) all share the same topology: an extracellular ligand-binding domain, a single transmembrane region and a cytoplasmic kinase domain. However they can be classified into at least five groups. The class III RTK's are characterized by the presence of five to seven immunoglobulin-like domains [2] in their extracellular section.

Their kinase domain differs from that of other RTK's by the insertion of a stretch of 70 to 100

hydrophilic residues in the middle of this domain. The receptors currently known to belong to class III are: - Platelet-derived growth factor receptor (PDGF-R). PDGF-R exists as a homo- or heterodimer of two related chains: alpha and beta [3]. - Macrophage colony stimulating factor receptor (CSF-1-R) (also known as the *fms* oncogene). - Stem cell factor (mast cell growth factor) receptor (also known as the *kit* oncogene). - Vascular endothelial growth factor (VEGF) receptors Flt-1 and Flk-1/KDR [4]. - Fl cytokine receptor Flk-2/Flt-3 [5]. - The putative receptor Flt-4 [7]. a signature pattern was developed for this class of RTK's which is based on a conserved region in the kinase domain.

10 Consensus pattern: G-x-H-x-N-[LIVM]-V-N-L-L-G-A-C-T-

- [1] Yarden Y., Ullrich A. *Annu. Rev. Biochem.* 57:443-478(1988).
- [2] Hunkapiller T., Hood L. *Adv. Immunol.* 44:1-63(1989).
- [3] Lee K.-H., Bowen-Pope D.F., Reed R.R. *Mol. Cell. Biol.* 10:2237-2246(1990).
- [4] Terman B.I., Dougher-Vermazen M., Carrion M.E., Dimitrov D., Armellino D.C., Gospodarowicz D., Boehlen P. *Biochem. Biophys. Res. Commun.* 187:1579-1586(1992).
- [5] Lyman S.D., James L., Vanden Bos T., de Vries P., Brasel K., Gliniak B., Hollingsworth L.T., Picha K.S., McKenna H.J., Splett R.R. *Cell* 75:1157-1167(1993).
- [6] Galland F., Karamysheva A., Pebusque M.J., Borg J.P., Rottapel R., Dubreuil P., Rosnet O., Birnbaum D. *Oncogene* 8:1233-1240(1993).

Receptor tyrosine kinase class V signatures

A number of growth factors stimulate mitogenesis by interacting with a family of cell surface receptors which possess an intrinsic, ligand-sensitive, protein tyrosine kinase activity [1].

- 25 These receptor tyrosine kinases (RTK) all share the same topology: an extracellular ligand-binding domain, a single transmembrane region and a cytoplasmic kinase domain. However they can be classified into at least five groups on the basis of sequence similarities. The extracellular domain of class V RTK's consist of a region of about 300 amino acids, amongst which 16 conserved cysteines probably involved in disulfide bonds; this region is followed
- 30 by two copies of a fibronectin type III domain. The ligands for these receptors are proteins of about 200 to 300 residues collectively known as Ephrins. The receptors currently known to belong to class V are [2,3,E1]: - EPHA1 (Eph-1; Esk). - EPHA2 (Eck; Mpk-5; Sek-2). - EPHA3 (Etk-1; Hek; Mek4; Tyro4; Rek4; Cek4). - EPHA4 (Sek; Hek8; Mpk-3; Cek8). -

425

EPHA5 (Ehk-1; Hek7; Bsk; Cek7). - EPHA6 (Ehk-2). - EPHA7 (Ehk-3; Hek11; Mdk-1; Ebk). - EPHA8 (Eek). - EPHB1 (Eph-2; Elk; Net). - EPHB2 (Eph-3; Hek5; Drt; Erk; Nuk; Sek-3; Cek5; Qek5). - EPHB3 (Hek-2; Mdk-5). - EPHB4 (Htk; Mdk-2; Myk-1). - EPHB5 (Cek9). The EPHA subtype receptors bind to GPI-anchored ephrins while the EPHB subtype
 5 receptors bind to type-I membrane ephrins. Two signature patterns were developed for this class of RTK's, which each include some of the conserved cysteine residues.

Consensus pattern: F-x-[DN]-x-[GAW]-[GA]-C-[LIVM]-[SA]-[LIVM](2)-[SA]-[LV]-
 10 [KRHQ]-[LIVA]-x(3)-[KR]-C-[PSAW] [The two C's are probably involved in disulfide bonds]

Consensus pattern: C-x(2)-[DE]-G-[DEQ]-W-x(2,3)-[PAQ]-[LIVMT]-[GT]-x-C-x-C- x(2)-
 G-[HFY]-[EQ] [The three C's are probably involved in disulfide bonds]

[1] Yarden Y., Ullrich A. Annu. Rev. Biochem. 57:443-478(1988).

[2] Sajjadi F.G., Pasquale E.B., Subramani S. New Biol. 3:769-778(1991).

[3] Wicks I.P., Wilkinson D., Salvaris E., Boyd A.W. Proc. Natl. Acad. Sci. U.S.A. 89:1611-
 1615(1992).

20 459. Protein kinase C terminal domain

460. Plant thionins signature

Thionins are small, basic, plant proteins generally toxic to animal cells [1]. They seem to exert
 25 their toxic effect at the level of the cell membrane but their exact function is not known. They consist of a polypeptide chain of forty five to fifty amino acids with three to four internal disulfide bonds. They are found in seeds but also in the cell wall of leaves [2]. Thionins are processed from larger precursor proteins [3]. Crambin [4], a hydrophobic plant seed protein, also belongs to this family. The pattern to detect this family of proteins includes three of the
 30 six cysteine residues involved in disulfide bonds. +-----+ | +-----

-----+ ||| | xxCCxxxxxxxxxxxxCxxxxxxxxCxxxCxxCxxxxxCxxxxxxxx

***** ||| +-----+'C': conserved cysteine involved in a disulfide bond.'*':

position of the pattern.

Consensus pattern: C-C-x(5)-R-x(2)-[FY]-x(2)-C [The three C's are involved in disulfide bonds] The proteins from the gamma-thionin family are not related to the above proteins and are described in a separate section.

5

[1] Vernon L.P., Evett G.E., Zeikus R.D., Gray W.R. Arch. Biochem. Biophys. 238:18-29(1985).

[2] Bohlmann H., Clausen S., Behnke S., Giese H., Hiller C., Reimann-Phillip U., Schrader G., Barkholt V., Apel K. EMBO J. 7:1559-1565(1988).

10 [3] Bohlmann H., Apel K. Mol. Gen. Genet. 207:446-454(1987).

[4] Teeter M.M., Mazer J.A., L'Italien J.J. Biochemistry 20:5437-5443(1981).

461. Polyprenyl synthetases signatures

15

A variety of isoprenoid compounds are synthesized by various organisms. For example in eukaryotes the isoprenoid biosynthetic pathway is responsible for the synthesis of a variety of end products including cholesterol, dolichol, ubiquinone or coenzyme Q. In bacteria this pathway leads to the synthesis of isopentenyl tRNA, isoprenoid quinones, and sugar carrier lipids. Among the enzymes that participate in that pathway, are a number of polyprenyl synthetase enzymes which catalyze a 1'4-condensation between 5 carbon isoprene units.

20

Currently the sequence of some of these enzymes is known: - Eukaryotic farnesyl pyrophosphate synthetase (FPP synthetase) (EC 2.5.1.1 / EC 2.5.1.10) which catalyzes the sequential condensation of isopentenyl pyrophosphate (IPP) with dimethylallyl pyrophosphate (DMAPP), and then with the resultant geranyl pyrophosphate to form farnesyl pyrophosphate. FPP synthetase is a cytoplasmic dimeric enzyme. - Prokaryotic farnesyl pyrophosphate synthetase (gene ispA). - Prokaryotic octaprenyl diphosphate synthase (gene ispB). - Prokaryotic heptaprenyl diphosphate synthase (EC 2.5.1.30). - Eukaryotic geranylgeranyl pyrophosphate synthetase (GGPP synthetase) (EC 2.5.1.1 / EC 2.5.1.10 / EC 2.5.1.29) which catalyzes the sequential addition of the three molecules of IPP onto DMAPP

25

30 to form geranylgeranyl pyrophosphate. In plants GGPP synthase is a chloroplast enzyme involved in the biosynthesis of terpenoids; in fungi, such as *Neurospora crassa* (gene al-3), this enzyme is involved in the biosynthesis of carotenoids. - Prokaryotic GGPP synthetase, which are involved in the biosynthesis of carotenoids (gene crtE). Such an enzyme is also

encoded in the cyanelle genome of *Cyanophora paradoxa*. - Eukaryotic hexaprenyl pyrophosphate synthetase, which is involved in the biosynthesis of coenzyme Q and which catalyzes the formation of all trans- polyprenyl pyrophosphates generally ranging in length of between 6 and 10 isoprene units depending on the species. HP synthetase is a mitochondrial membrane-associated enzyme. It has been shown [1 to 5] that all the above enzymes share some regions of sequence similarity. Two of these regions are rich in aspartic-acid residues and could be involved in the catalytic mechanism and/or the binding of the substrates. signature patterns were developed for both regions. Possible additional members of this family of proteins are: - *Bacillus subtilis* spore germination protein C3 (gene gerC3). Both proteins are most probably also enzymes involved in isoprenoid metabolism [6].

Consensus pattern: [LIVM](2)-x-D-D-x(2,4)-D-x(4)-R-R-[GH]-

Consensus pattern: [LIVMFY]-G-x(2)-[FYL]-Q-[LIVM]-x-D-D-[LIVMFY]-x-[DNG]

[1] Ashby M.N., Edwards P.A. J. Biol. Chem. 265:13157-13164(1990).

[2] Fujisaki S., Hara H., Nishimura Y., Horiuchi K., Nishino T. J. Biochem. 108:995-1000(1990).

[3] Carattoli A., Romano N., Ballario P., Morelli G., Macino G. J. Biol. Chem. 266:5854-5859(1991).

[4] Kuntz M., Roemer S., Suire C., Hugueney P., Weil J.H., Schantz R., Camara B. Plant J. 2:25-34(1992).

[5] Math S.K., Hearst J.E., Poulter C.D. Proc. Natl. Acad. Sci. U.S.A. 89:6761-6764(1992).

[6] Bairoch A. Unpublished observations (1993).

462. Potato inhibitor I family signature

The potato inhibitor I family is one of the numerous families of serine proteinase inhibitors. Members of this protein family are found in plants; in the seeds of barley or beans [1,2,3], and in potato or tomato leaves where they accumulate in response to mechanical damage [4,5]. An inhibitor belonging to this family is also found in leech [6]. It is interesting to note that, currently, this is the only proteinase inhibitor family to be found both in plant and animal kingdoms. Structurally these inhibitors are small (60 to 90 residues) and in contrast with other families of protease inhibitors, they lack disulfide bonds. They have a single inhibitory

site. The consensus pattern includes three out of the four residues conserved in all members of this family and is located in the N-terminal half.

Consensus pattern: [FYW]-P-[EQH]-[LIV](2)-G-x(2)-[STAGV]-x(2)-A- Barley subtilisin-
 5 chymotrypsin inhibitor-2b has Glu instead of Gly. There is a trypsin inhibitor from the cucurbitaceae *Momordica charantia* [7], which is said to belong to the potato inhibitor I family but which shows only a very weak similarity with the other members of this family.

[1] Svendsen I., Hejgaard J., Chavan J.K. Carlsberg Res. Commun. 49:493-502(1984).

10 [2] Svendsen I., Boisen S., Hejgaard J. Carlsberg Res. Commun. 47:45-53(1982).

[3] Nozawa H., Yamagata H., Aizono Y., Yoshikawa M., Iwasaki T. J. Biochem. 106:1003-1008(1989).

[4] Cleveland T.E., Thornburg R.W., Ryan C.A. Plant Mol. Biol. 8:199-207(1987).

[5] Lee J.S., Brown W.E., Graham J.S., Pearce G., Fox E.A., Dreher T.W., Ahern K.G., Pearson G.D., Ryan C.A. Proc. Natl. Acad. Sci. U.S.A. 83:7277-7281(1986).

[6] Seemuller U., Eulitz M., Fritz H., Strobl A. Hoppe-Seyler's Z. Physiol. Chem. 361:1841-1846(1980).

[7] Zeng F.-Y., Qian R.-Q., Wang Y. FEBS Lett. 234:35-38(1988).

463. (pp binding) Phosphopantetheine attachment site

Phosphopantetheine (or pantetheine 4' phosphate) is the prosthetic group of acyl carrier proteins (ACP) in some multienzyme complexes where it serves as a 'swinging arm' for the attachment of activated fatty acid and amino-acid groups [1]. Phosphopantetheine is attached
 25 to a serine residue in these proteins [2]. ACP proteins or domains have been found in various enzyme systems which are listed below (references are only provided for recently determined sequences). - Fatty acid synthetase (FAS), which catalyzes the formation of long-chain fatty acids from acetyl-CoA, malonyl-CoA and NADPH. Bacterial and plant chloroplast FAS are composed of eight separate subunits which correspond to the different enzymatic activities;
 30 ACP is one of these polypeptides. Fungal FAS consists of two multifunctional proteins, FAS1 and FAS2; the ACP domain is located in the N-terminal section of FAS2. Vertebrate FAS consists of a single multifunctional enzyme; the ACP domain is located between the beta-ketoacyl reductase domain and the C-terminal thioesterase domain [3]. - Polyketide

antibiotics synthase enzyme systems. Polyketides are secondary metabolites produced from simple fatty acids, by microorganisms and plants. ACP is one of the polypeptidic components involved in the biosynthesis of *Streptomyces* polyketide antibiotics actinorhodin, curamycin, granatacin, monensin, oxytetracycline and tetracenomycin C. - *Bacillus subtilis* putative polyketide synthases pksK, pksL and pksM which respectively contain three, five and one ACP domains. - The multifunctional 6-methylsalicylic acid synthase (MSAS) from *Penicillium patulum*. This is a multifunctional enzyme involved in the biosynthesis of a polyketide antibiotic and which contains an ACP domain in the C-terminal extremity. - Multifunctional mycocerosic acid synthase (gene mas) from *Mycobacterium bovis*. - Gramicidin S synthetase I (gene grsA) from *Bacillus brevis*. This enzyme catalyzes the first step in the biosynthesis of the cyclic antibiotic gramicidin S. - Tyrocidine synthetase I (gene tycA) from *Bacillus brevis*. The reaction carried out by tycA is identical to that catalyzed by grsA - Gramicidin S synthetase II (gene grsB) from *Bacillus brevis*. This enzyme is a multifunctional protein that activates and polymerizes proline, valine, ornithine and leucine. GrsB contains four ACP domains. - Erythronolide synthase proteins 1, 2 and 3 from *Saccharopolyspora erythraea* which is involved in the biosynthesis of the polyketide antibiotic erythromycin. Each of these proteins contain two ACP domains. - Conidial green pigment synthase from *Aspergillus nidulans*. - ACV synthetase from various fungi. This enzyme catalyzes the first step in the biosynthesis of penicillin and cephalosporin. It contains three ACP domains. - Enterobactin synthetase component F (gene entF) from *Escherichia coli*. This enzyme is involved in the ATP-dependent activation of serine during enterobactin (enterochelin) biosynthesis. - Cyclic peptide antibiotic surfactin synthase subunits 1, 2 and 3 from *Bacillus subtilis*. Subunits 1 and 2 contains three related domains while subunit 3 only contains a single domain. - HC-toxin synthetase (gene HTS1) from *Cochliobolus carbonum*. This enzyme synthesizes HC-toxin, a cyclic tetrapeptide. HTS1 contains four ACP domains. - Fungal mitochondrial ACP [9], which is part of the respiratory chain NADH dehydrogenase (complex I). - *Rhizobium* nodulation protein nodF, which probably acts as an ACP in the synthesis of the nodulation Nod factor fatty acyl chain. The sequence around the phosphopantetheine attachment site is conserved in all these proteins and can be used as a signature pattern. A profile was also developed that spans the complete ACP-like domain.

Consensus pattern: [DEQGSTALMKRH]-[LIVMFYSTAC]-[GNQ]-[LIVMFYAG]-
[DNEKHS]-S-[LIVMST]-{PCFY}-[STAGCPQLIVMF]-[LIVMATN]-

[DENQGTAKRHLN]-[LIVMWSTA]-[LIVGSTACR]-x(2)-[LIVMFA] [S is the pantetheine attachment site]

[1] Concise Encyclopedia Biochemistry, Second Edition, Walter de Gruyter, Berlin New-York (1988).

[2] Pugh E.L., Wakil S.J. J. Biol. Chem. 240:4727-4733(1965).

[3] Witkowski A., Rangan V.S., Randhawa Z.I., Amy C.M., Smith S. Eur. J. Biochem. 198:571-579(1991).

[6] Scotti C., Piatti M., Cuzzoni A., Perani P., Tognoni A., Grandi G., Galizzi A., Albertini A.M. Gene 130:65-71(1993).

[9] Sackmann U., Zensen R., Rohlen D., Jahnke U., Weiss H. Eur. J. Biochem. 200:463-469(1991).

464. (Prenyltrans) Terpene synthases signature

The following enzymes catalyze mechanistically related reactions which involve the highly complex cyclic rearrangement of squalene or its 2,3 oxide: - Lanosterol synthase (EC 5.4.99.7) (oxidosqualene--lanosterol cyclase), which catalyzes the cyclization of (S)-2,3-epoxysqualene to lanosterol, the initial precursor of cholesterol, steroid hormones and vitamin D in vertebrates and of ergosterol in fungi (gene ERG7). - Cycloartenol synthase (EC 5.4.99.8) (2,3-epoxysqualene--cycloartenol cyclase), a plant enzyme that catalyzes the cyclization of (S)-2,3-epoxysqualene to cycloartenol. - Hopene synthase (EC 5.4.99.-) (squalene--hopene cyclase), a bacterial enzyme that catalyzes the cyclization of squalene into hopene, a key step in hopanoid (triterpenoid) metabolism. These enzymes are evolutionary related [1] proteins of about 70 to 85 Kd. As a signature pattern, a highly conserved region was selected which is rich in aromatic residues and which is located in the C-terminal section.

Consensus pattern: [DE]-G-S-W-x-G-x-W-[GA]-[LIVM]-x-[FY]-x-Y-[GA]

[1] Corey E.J., Matsuda S.P.T., Bartel B. Proc. Natl. Acad. Sci. U.S.A. 90:11628-11632(1993).

465. Prion protein signatures

Prion protein (PrP) [1,2,3] is a small glycoprotein found in high quantity in the brains of humans or animals infected with a number of degenerative neurological diseases such as Kuru, Creutzfeldt-Jacob disease (CJD), scrapie or bovine spongiform encephalopathy (BSE).

PrP is encoded in the host genome and expressed both in normal and infected cells. It has a tendency to aggregate yielding polymers called rods. Structurally, PrP is a protein consisting of a signal peptide, followed by an N-terminal domain that contains tandem repeats of a short motif (PHGGGWGQ in mammals, PHNPGY in chicken), itself followed by a highly conserved domain. It comes a C-terminal hydrophobic domain post-translationally removed when PrP is attached to the extracellular side of the cell membrane by a GPI-anchor. The structure of PrP is shown in the following schematic representation: +---+-----+---

*****-----*****-----+---+ |Sig| Tandem repeats | C C S | | +---+-----+---
-----|-----|-----|+---+ +-----+ | GPI-C': conserved cysteine involved in a disulfide bond. '*': position of the patterns. As signature pattern for PrP, a perfectly conserved alanine- and glycine-rich region of 16 residues was selected as well as a region centered on the second cysteine involved in the disulfide bond.

Consensus pattern: A-G-A-A-A-A-G-A-V-V-G-G-L-G-G-Y-

Consensus pattern: E-x-[ED]-x-K-[LIVM](2)-x-[KR]-[LIVM](2)-x-[QE]-M-C-x(2)-Q-Y [C is involved in a disulfide bond]

[1] Stahl N., Prusiner S.B. FASEB J. 5:2799-2807(1991).

[2] Brunori M., Chiara Silvestrini M., Pocchiari M. Trends Biochem. Sci. 13:309-313(1988).

[3] Prusiner S.B. Annu. Rev. Microbiol. 43:345-374(1989).

466. Cyclophilin-type peptidyl-prolyl cis-trans isomerase signature and profile (pro isomerase)

Cyclophilin [1] is the major high-affinity binding protein in vertebrates for the immunosuppressive drug cyclosporin A (CSA). It exhibits a peptidyl- prolyl cis-trans isomerase activity (EC 5.2.1.8) (PPIase or rotamase). PPIase is an enzyme that accelerates protein folding by catalyzing the cis-trans isomerization of proline imidic peptide bonds in oligopeptides [2]. It is probable that CSA mediates some of its effects via an inhibitory action

on PPIase. Cyclophilin is a cytosolic protein which belongs to a family [3,4,5] that also includes the following isozymes: - Cyclophilin B (or S-cyclophilin), a PPIase which is retained in an endoplasmic reticulum compartment. - Cyclophilin C, a cytoplasmic PPIase. - Mitochondrial matrix cyclophilin (cyp3). - A PPIase which seems specific for the folding of rhodopsin and is an integral membrane protein anchored by a C-terminal transmembrane region. This protein was first characterized in *Drosophila* (gene *ninaA*). - Bacterial periplasmic PPIase (gene *ppiA*). - Bacterial cytosolic PPIase (gene *ppiB*). - Natural-killer cell cyclophilin-related protein. This large protein (about 160 Kd) is a component of a putative tumor-recognition complex involved in the function of NK cells. It contains a cyclophilin-type PPIase domain. - Mammalian nucleoporin Nup358 [6], a nuclear pore complex protein of 358 Kd that contains a C-terminal cyclophilin-type PPIase domain. - Yeast hypothetical protein YJR032w. - Fission yeast hypothetical protein SpAC21E11.05c. - *Caenorhabditis elegans* hypothetical protein T27D1.1. The sequences of the different forms of cyclophilin-type PPIases are well conserved. As a signature pattern, a conserved region was selected in the central part of these enzymes.

Consensus pattern: [FY]-x(2)-[STCNLV]-x-F-H-[RH]-[LIVMN]-[LIVM]-x(2)-F-[LIVM]-x-Q-[AG]-G- FKBP's, a family of proteins that bind the immunosuppressive drug FK506, are also PPIases, but their sequence is not at all related to that of cyclophilin.

[1] Stamnes M.A., Rutherford S.L., Zuker C.S. *Trends Cell Biol.* 2:272-276(1992).

[2] Fischer G., Schmid F.X. *Biochemistry* 29:2205-2212(1990).

[3] Trandinh C.C., Pao G.M., Saier M.H. Jr. *FASEB J.* 6:3410-3420(1992).

[4] Galat A. *Eur. J. Biochem.* 216:689-707(1993).

[5] Hacker J., Fischer G. *Mol. Microbiol.* 10:445-456(1993).

[6] Wu J., Matunis M.J., Kraemer D., Blobel G., Coutavas E. *J. Biol. Chem.* 270:14209-14213(1995).

467. Profilin signature

Profilin [1,2] is a small eukaryotic protein that binds to monomeric actin(G-actin) in a 1:1 ratio thus preventing the polymerization of actin into filaments (F-actin). It can also, in certain circumstance promotes actin polymerization. Profilin also binds to

433

polyphosphoinositides such as PIP2. Overall sequence similarity among profilin from organisms which belong to different phyla (ranging from fungi to mammals) is low, but the N-terminal region is relatively well conserved. That region is thought to be involved in the binding to actin. The signature pattern for profilin is based on conserved residues at the N-terminal extremity. A protein structurally similar to profilin is present in the genome of variola and vaccinia viruses (gene A42R).

Consensus pattern: <x(0,1)-[STA]-x(0,1)-W-[DENQH]-x-[YI]-x-[DEQ]

- [1] Haarer B.K., Brown S.S. Cell Motil. Cytoskeleton 17:71-74(1990).
[2] Sohn R.H., Goldschmidt-Clermont P. BioEssays 16:465-472(1994).

468. Protamine P1 signature

Protamines are small, highly basic proteins, that substitute for histones in sperm chromatin during the haploid phase of spermatogenesis. They pack sperm DNA into a highly condensed, stable and inactive complex. There are two different types of mammalian protamine, called P1 and P2. P1 has been found in all species studied, while P2 is sometimes absent. There seems to be a single type of avian protamine whose sequence is closely related to that of mammalian P1 [1]. As a signature for this family of proteins, a conserved region was selected at the N-terminal extremity of the sequence.

Consensus pattern: [AV]-R-[NFY]-R-x(2,3)-[ST]-x-S-x-S-

- [1] Oliva R., Goren R., Dixon G.H. J. Biol. Chem. 264:17627-17630(1989).

469. Sperm histone P2 (protamine P2)

This protein also known as protamine P2 can substitute for histones in the chromatin of sperm. The alignment contains both the sequence of the mature P2 protein and its propeptide.

470. Proteasome A-type subunits signature

The proteasome (or macropain) (EC 3.4.99.46) [1 to 5,E1] is an eukaryotic and archaeobacterial multicatalytic proteinase complex that seems to be involved in an ATP/ubiquitin-dependent nonlysosomal proteolytic pathway. In eukaryotes the proteasome is composed of about 28 distinct subunits which form a highly ordered ring-shaped structure (20S ring) of about 700 Kd. Most proteasome subunits can be classified, on the basis on sequence similarities into two groups, A and B. Subunits that belong to the A-type group are proteins of from 210 to 290 amino acids that share a number of conserved sequence regions. Subunits that are known to belong to this family are listed below. - Vertebrate subunits C2 (nu), C3, C8, C9, iota and zeta. - Drosophila PROS-25, PROS-28.1, PROS-29 and PROS-35. - Yeast C1 (PRS1), C5 (PRS3), C7-alpha (Y8) (PRS2), Y7, Y13, PRE5, PRE6 and PUP2. - Arabidopsis thaliana subunits alpha and PSM30. - Thermoplasma acidophilum alpha-subunit. In this archaeobacteria the proteasome is composed of only two different subunits. As a signature pattern for proteasome A-type subunits the best conserved region was selected, which is located in the N-terminal part of these proteins.

Consensus pattern: [FY]-x(4)-[STNV]-x-[FYW]-S-P-x-G-[RKH]-x(2)-Q-[LIVM]-[DE]-Y-[SAD]-x(2)-[SAG]-. These proteins belong to family T1 in the classification of peptidases [6,E2].

- [1] Rivett A.J. Biochem. J. 291:1-10(1993).
- [2] Rivett A.J. Arch. Biochem. Biophys. 268:1-8(1989).
- [3] Goldberg A.L., Rock K.L Nature 357:375-379(1992).
- [4] Wilk S. Enzyme Protein 47:187-188(1993).
- [5] Hilt W., Wolf D.H. Trends Biochem. Sci. 21:96-102(1996).
- [6] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

Proteasome B-type subunits signature

The proteasome (or macropain) (EC 3.4.99.46) [1 to 5,E1] is an eukaryotic and archaeobacterial multicatalytic proteinase complex that seems to be involved in an ATP/ubiquitin-dependent nonlysosomal proteolytic pathway. In eukaryotes the proteasome is composed of about 28 distinct subunits which form a highly ordered ring-shaped structure (20S ring) of about 700 Kd. Most proteasome subunits can be classified, on the basis on sequence similarities into two groups, A and B. Subunits that belong to the B-type group are

proteins of from 190 to 290 amino acids that share a number of conserved sequence regions.

Subunits that are known to belong to this family are listed below. - Vertebrate subunits C5, beta, delta, epsilon, theta (C10-II), LMP2/RING12, C13 (LMP7/RING10), C7-I and MECL-

1. - Yeast PRE1, PRE2 (PRG1), PRE3, PRE4, PRS3, PUP1 and PUP3. - Drosophila

5 L(3)73AI. - Fission yeast pts1. - Thermoplasma acidophilum beta-subunit. In this archaeobacteria the proteasome is composed of only two different subunits. As a signature pattern for proteasome B-type subunits the best conserved region was selected, which is located in the N-terminal part of these proteins.

10 Consensus pattern: [LIVMA]-[GSA]-[LIVMF]-x-[FYLVGAC]-x(2)-[GSACFY]-[LIVMSTAC](3)-[GAC]-[GSTACV]-[DES]-x(15)-[RK]-x(12,13)-G-x(2)-[GSTA]-D-. These proteins belong to family T1 in the classification of peptidases [6,E2].

[1] Rivett A.J. Biochem. J. 291:1-10(1993).

[2] Rivett A.J. Arch. Biochem. Biophys. 268:1-8(1989).

[3] Goldberg A.L., Rock K.L Nature 357:375-379(1992).

[4] Wilk S. Enzyme Protein 47:187-188(1993).

[5] Hilt W., Wolf D.H. Trends Biochem. Sci. 21:96-102(1996).

[6] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

471. (pyr redox) Pyridine nucleotide-disulphide oxidoreductases class-I active site

The pyridine nucleotide-disulphide oxidoreductases are FAD flavoproteins which contains a pair of redox-active cysteines involved in the transfer of reducing equivalents from the FAD

25 cofactor to the substrate. On the basis of sequence and structural similarities [1] these enzymes can be classified into two categories. The first category groups together the following enzymes [2 to 6]: - Glutathione reductase (EC 1.6.4.2) (GR). - Higher eukaryotes thioredoxin reductase (EC 1.6.4.5). - Trypanothione reductase (EC 1.6.4.8). - Lipoamide dehydrogenase (EC 1.8.1.4), the E3 component of alpha-ketoacid dehydrogenase complexes.

30 - Mercuric reductase (EC 1.16.1.1). The sequence around the two cysteines involved in the redox-active disulfide bond is conserved and can be used as a signature pattern.

Consensus pattern: G-G-x-C-[LIVA]-x(2)-G-C-[LIVM]-P [The two C's form the active site disulfide bond]. In positions 6 and 7 of the pattern all known sequences have Asn-(Val/ Ile) with the exception of GR from plant chloroplasts and from cyanobacteria which have Ile-Arg [7].

5

[1] Kurlyan J., Krishna T.S.R., Wong L., Guenther B., Pahler A., Williams C.H. Jr., Model P. Nature 352:172-174(1991).

[2] Rice D.W., Schulz G.E., Guest J.R. J. Mol. Biol. 174:483-496(1984).

[3] Brown N.L. Trends Biochem. Sci. 10:400-402(1985).

10 [4] Carothers D.J., Pons G., Patel M.S. Arch. Biochem. Biophys. 268:409-425(1989).

[5] Walsh C.T., Bradley M., Nadeau K. Trends Biochem. Sci. 16:305-309(1991).

[6] Gasdaska P.Y., Gasdaska J.R., Cochran S., Powis G. FEBS Lett. 373:5-9(1995).

[7] Creissen G., Edwards E.A., Enard C., Wellburn A., Mullineaux P. Plant J. 2:129-131(1991).

15

472. (pyridoxal deC) DDC / GAD / HDC / TyrDC pyridoxal-phosphate attachment site (pyridoxal deC)

Three different enzymes - all pyridoxal-dependent decarboxylases - seem to share regions of sequence similarity [1,2,3,4], especially in the vicinity of the lysine residue which serves as the attachment site for the pyridoxal-phosphate (PLP) group. These enzymes are: - Glutamate decarboxylase (EC 4.1.1.15) (GAD). Catalyzes the decarboxylation of glutamate into the neurotransmitter GABA (4-aminobutanoate). - Histidine decarboxylase (EC 4.1.1.22) (HDC). Catalyzes the decarboxylation of histidine to histamine. There are two completely unrelated

25 types of HDC: those that use PLP as a cofactor (found in Gram-negative bacteria and mammals), and those that contain a covalently bound pyruvoyl residue (found in Gram-positive bacteria). - Aromatic-L-amino-acid decarboxylase (EC 4.1.1.28) (DDC), also known as L-dopa decarboxylase or tryptophan decarboxylase. DDC catalyzes the decarboxylation of tryptophan to tryptamine. It also acts on 5-hydroxy- tryptophan and dihydroxyphenylalanine

30 (L-dopa). - Tyrosine decarboxylase (EC 4.1.1.25) (TyrDC) which converts tyrosine into tyramine, a precursor of isoquinoline alkaloids and various amides. These enzymes are collectively known as group II decarboxylases [3,4].

437

Consensus pattern: S-[LIVMFYW]-x(5)-K-[LIVMFYWG](2)-x(3)-[LIVMFYW]-x-[CA]-
x(2)-[LIVMFYWQ]-x(2)-[RK] [K is the pyridoxal-P attachment site]

[1] Jackson F.R. J. Mol. Evol. 31:325-329(1990).

5 [2] Joseph D.R., Sullivan P., Wang Y.-M., Kozak C., Fenstermacher D.A., Behrendsen M.E.,
Zahnow C.A. Proc. Natl. Acad. Sci. U.S.A. 87:733-737(1990).

[3] Sandmeier E., Hale T.I., Christen P. Eur. J. Biochem. 221:997-1002(1994).

[4] Ishii S., Mizuguchi H., Nishino J., Hayashi H., Kagamiyama H. J. Biochem. 120:369-
376(1996).

10

473. Regulator of chromosome condensation (RCC1) signatures (RCC1)

The regulator of chromosome condensation (RCC1) [1] is a eukaryotic protein which binds to
chromatin and interacts with ran, a nuclear GTP-binding protein, to promote the loss of
bound GDP and the uptake of fresh GTP, thus acting as a guanine-nucleotide dissociation
stimulator (GDS)[2]. The interaction of RCC1 with ran probably plays an important role in
the regulation of gene expression. RCC1, known as PRP20 or SRM1 in yeast, pim1 in fission
yeast and BJI in Drosophila, is a protein that contains seven tandem repeats of a domain of
about 50 to 60 amino acids. As shown in the following schematic representation, the repeats
make up the major part of the length of the protein. Outside the repeat region, there is just a
small N-terminal domain of about 40 to 50 residues and, in the Drosophila protein only, a C-
terminal domain of about 130 residues. +----+-----+-----+-----+-----+-----+-----+-----+-----+
--+-+-----+ |N-t.|Rpt. 1 |Rpt. 2 |Rpt. 3 |Rpt. 4 |Rpt. 5 |Rpt. 6 |Rpt. 7 | C-terminal | +-----+
----+-----+-----+-----+-----+-----+-----+-----+-----+ In Drosophila two signature

25 patterns for RCC1 were developed. The first is found in the N- terminal part of the second
repeat; this is the most conserved part of RCC1. The second is derived from conserved
positions in the C-terminal part of each repeat and detects up to five copies of the repeated
domain. The RCC1-type of repeat is also found in the X-linked retinitis pigmentosa GTPase
regulator [3].

30

Consensus pattern: G-x-N-D-x(2)-[AV]-L-G-R-x-T-

Consensus pattern: [LIVMFA]-[STAGC](2)-G-x(2)-H-[STAGLI]-[LIVMFA]-x-[LIVM]-

- [1] Dasso M. Trends Biochem. Sci. 18:96-101(1993).
 [2] Boguski M.S., McCormick F. Nature 366:643-654(1993).
 [3] Roepman R., Van Duijnhoven G., Rosenberg T., Pinckers A.J.L.G., Bleeker-
 Wagemakers L.M., Bergen A.A.B., Post J., Beck A., Reinhardt R., Ropers H.-H., Cremers F.,
 5 Berger W. Hum. Mol. Genet. 5:1035-1041(1996).

474. RNA 3'-terminal phosphate cyclase signature (RCT)

RNA 3'-terminal phosphate cyclase (EC 6.5.1.4) [1,2] catalyzes the conversion of 3'-
 10 phosphate to a 2',3'-cyclic phosphodiester at the end of RNA. The biological role of this
 enzyme is unknown but it is likely to function in some aspects of cellular RNA processing.
 The reaction catalyzed by the enzyme occurs in three steps: 1) adenylation of the enzyme by
 ATP; 2) the enzyme acts on RNA-3'terminal phosphate to produce RNA-3'terminal
 15 diphosphate adenylate; 3) Release of AMP and cyclisation by a non catalytic nucleophilic
 attack by the adjacent 2'hydroxyl on the phosphorus in the diester linkage. This enzyme,
 which has been characterized in human (where there seems to be at least three isozymes) and
 Escherichia coli (gene rtCA), seems to be taxonomically widespread. It is found in insects,
 plants, fungi (gene RTC1 in yeast) and in archeabacteria. RNA cyclase is a protein of from 36
 to 42 Kd. The best conserved region, which is used as a signature pattern, is a glycine-rich
 20 stretch of residues located in the central part of the sequence and which is reminiscent of
 various ATP, GTP or AMP glycine-rich loops. In this context, the conserved Arg (His in the
 E.coli enzyme) could be the AMP-binding residue.

Consensus pattern: [RH]-G-x(2)-P-x-G(3)-x-[LIV]-

- 25 [1] Genschik P., Billy E., Swianiewicz M., Filipowicz W. EMBO J. 16:2955-2967(1997).
 [2] Filipowicz W., Vincente O. Meth. Enzymol. 181:499-510(1990).

30 475. REV protein (anti-repression trans-activator protein)

476. Prokaryotic-type class I peptide chain release factors signature (RF-1)

Peptide chain release factors (RFs) are required for the termination of protein biosynthesis [1]. At present two classes of RFs can be distinguished. Class I RFs bind to ribosomes that have encountered a stop codon at their decoding site and induce release of the nascent polypeptide. Class II RFs are GTP-binding proteins that interact with class I RFs and enhance class I RF activity. In prokaryotes there are two class I RFs that act in a codon specific manner[2]: RF-1 (gene prfA) mediates UAA and UAG-dependent termination while RF-2(gene prfB) mediates UAA and UGA-dependent termination. RF-1 and RF-2 are structurally and evolutionary related proteins which have been shown [3] to make up a family that also contains the following proteins: - Fungal MRF1, a mitochondrial RF (m-RF) which recognizes the UAA and UAG codons. - Escherichia coli RF-H, a protein of unknown function. - Escherichia coli hypothetical protein yaeJ and a close Pseudomonas putida homolog. A highly conserved region located in the central part of the 40 to 45 Kd RF-1/2 and m-RF and in the N-terminal of the 15 to 16Kd RF-H and yaeJ is used as a signature pattern.

Consensus pattern: [AR]-[STA]-x-G-x-G-G-Q-[HNGCS]-V-N-x(3)-[ST]-A-[IV]

Note that prokaryotic-type class I RFs display no significant sequence similarity to prokaryotic-type class II which belong to the family of GTP-binding elongation factors nor to eukaryotic class I or class II RFs.

[1] Tate W.P. , Poole E.S., Mannering S.M. Prog. Nucleic Acids. Res. Mol. Biol. 52:293-335(1996).

[2] Craigen W.J., Lee C.C., Caskey C.T. Mol. Microbiol. 4:861-865(1990).

[3] Pel H.J., Rep M., Grivell L.A. Nucleic Acids Res. 20:4423-4428(1992).

477. RIO1/ZK632.3/MJ0444 family signature

The following uncharacterized proteins are evolutionary related [1]: - Yeast protein RIO1. - Caenorhabditis elegans hypothetical protein ZK632.3. - Methanococcus jannaschii hypothetical protein MJ0444. - Thermoplasma acidophilum hypothetical protein if rpoA2 3'region. The eukaryotic members of this family are proteins of about 55 to 60 Kd, while the archebacterial ones are half that size. The central part of these proteins is highly conserved. The best conserved region is used as a signature pattern.

Consensus pattern: [LIVM]-V-H-[GA]-D-L-S-E-[FY]-N-x-[LIVM]

[1] Bairoch A. Unpublished observations (1997).

5

10

15

20

25

30

478. (RIP) Shiga/ricin ribosomal inactivating toxins active site signature. A number of bacterial and plant toxins act by inhibiting protein synthesis in eukaryotic cells. The toxins of the Shiga and ricin family inactivate 60S ribosomal subunits by an N-glycosidic cleavage which releases a specific adenine base from the sugar-phosphate backbone of 28S rRNA [1,2,3]. The toxins which are known to function in this manner are: - Shiga toxin from *Shigella dysenteriae* [4]. This toxin is composed of one copy of an enzymatically active A subunit and five copies of a B subunit responsible for binding the toxin complex to specific receptors on the target cell surface. - Shiga-like toxins (SLT) are a group of *Escherichia coli* toxins very similar in their structure and properties to Shiga toxin. The sequence of two types of these toxins, SLT-1 [5] and SLT-2 [6], is known. - Ricin, a potent toxin from castor bean seeds. Ricin consists of two glycosylated chains linked by a disulfide bond. The A chain is enzymatically active. The B chain is a lectin with a binding preference for galactosides. Both chains are encoded by a single polypeptidic precursor. Ricin is classified as a type-II ribosome-inactivating protein (RIP); other members of this family are agglutinin, also from castor bean, and abrin from the seeds of the bean *Abrus precatorius* [7]. - Single chain ribosome-inactivating proteins (type-I RIP) from plants. Examples of such proteins are: barley protein synthesis inhibitors I and II, mongolian snake-gourd trichosanthin, sponge gourd luffin-A and -B, garden four-o'clock MAP, common pokeberry PAP-S and soapwort saporin-6 [7]. All these toxins are structurally related. A conserved glutamic residue has been implicated [8] in the catalytic mechanism; it is located near a conserved arginine which also plays a role in catalysis [9]. The signature that has been developed for these proteins includes these catalytic residues.

Consensus pattern: [LIVMA]-x-[LIVMSTA](2)-x-E-[SAGV]-[STAL]-R-[FY]-[RKNQS]-x-[LIVM]-[EQS]-x(2)-[LIVMF] [E and R are active site residues]-

[1] Endo Y., Tsurugi K., Takeda Y., Ogasawara T., Igarashi K. *Eur. J. Biochem.* 171:45-50(1988). [2] May M.J., Hartley M.R., Roberts L.M., Krieg P.A., Osborn R.W., Lord J.M.

EMBO J. 8:301-308(1989).[3] Funatsu G., Islam M.R., Minami Y., Sung-Sil K., Kimura M. Biochimie 73:1157-1161(1991).[4] Strockbine N.A., Jackson M.P., Sung L.M., Holmes R.K., O'Brien A.D. J. Bacteriol. 170:1116-1122(1988).[5] Calderwood S.B., Auclair F., Donohue-Rolfe A., Keusch G.T., Mekalanos J.J. Proc. Natl. Acad. Sci. U.S.A. 84:4364-4368(1987).[6] Jackson M.P., Neill R.J., O'Brien A.D., Holmes R.K., Newland J.W. FEMS Microbiol. Lett. 44:109-114(1987).[7] Barbieri L., Battelli M.G., Stirpe F. Biochim. Biophys. Acta 1154:237-282(1993).[8] Hovde C.J., Calderwood S.B., Mekalanos J.J., Collier R.J. Proc. Natl. Acad. Sci. U.S.A. 85:2568-2572(1988).[9] Monzingo A.F., Collins E.J., Ernst S.R., Irvin J.D., Robertus J.D. J. Mol. Biol. 233:705-715(1993).

479. Bacterial RNA polymerase, alpha chain (RNA pol A bac)

Members of this family include alpha subunit from eubacteria and alpha subunits from chloroplasts. The alpha subunit of RNA polymerase consists of two independently folded domains, referred to as amino-terminal and carboxyl terminal domains. The amino terminal domain is involved in the interaction with the other subunits of the RNA polymerase. The carboxyl-terminal domain interacts with the DNA and activators. The amino acid sequence of the alpha subunit is conserved in prokaryotic and chloroplast RNA polymerases. There are three regions of particularly strong conservation, two in the amino-terminal and one in the carboxyl-terminal [3].

[1] Zhang G, Darst SA; Science 1998;281:262-266. [2] Jeon YH, Negishi T, Shirakawa M, Yamazaki T, Fujita N, Ishihama A, Kyogoku Y; Science 1995;270:1495-1497. [3] Ebright RH, Busby S; Curr Opin Genet Dev 1995;5:197-203. [4] Murakami K, Kimura M, Owens JT, Meares CF, Ishihama A; Proc Natl Acad Sci USA 1997;94:1709-1714.

480. RNA polymerase beta subunit (RNA pol B)

RNA polymerases catalyse the DNA dependent polymerisation of RNA. Prokaryotes contain a single RNA polymerase compared to three in eukaryotes (not including mitochondrial and chloroplast polymerases). Each RNA polymerase complex contains two related members of this family, in each case they are the two largest subunits.

[1] Falkenburg D, Dworniczak B, Faust DM, Bautz EK; J Mol Biol 1987;195:929-937.

481. RNA polymerases H / 23 Kd subunits signature

In eukaryotes, there are three different forms of DNA-dependent RNA polymerases (EC 2.7.7.6) transcribing different sets of genes. Each class of RNA polymerase is an assemblage of ten to twelve different polypeptides. In archaebacteria, there is generally a single form of RNA polymerase which also consist of an oligomeric assemblage of 10 to 13 polypeptides. Archaeobacterial subunit H (gene rpoH) [1,2] is a small protein of about 8.5 to 10 Kd, it is evolutionary related to the C-terminal part of a 23 Kd component shared by all three forms of eukaryotic RNA polymerases (gene RPB5 in yeast and POLR2E in mammals). As a signature pattern a conserved region was selected which is located at the N-terminal extremity of subunit H; this region contains two histidines that could play a role in the binding of a metal ion.

Consensus pattern: H-[NEI]-[LIVM]-V-P-x-H-x(2)-[LIVM]-x(2)-[DE]

[1] Klenk H.-P., Palm P., Lottspeich F., Zillig W. Proc. Natl. Acad. Sci. U.S.A. 89:407-410(1992).

[2] Thiru A., Hodach M., Eloranta J.J., Kostourou V., Weinzierl R.O., Matthews S.; J. Mol. Biol. 287:753-760(1999).

482. RNA polymerases K / 14 to 18 Kd subunits signature

In eukaryotes, there are three different forms of DNA-dependent RNA polymerases (EC 2.7.7.6) transcribing different sets of genes. Each class of RNA polymerase is an assemblage of ten to twelve different polypeptides. In archaebacteria, there is generally a single form of RNA polymerase which also consist of an oligomeric assemblage of 10 to 13 polypeptides. A component of 14 to 18 Kd shared by all three forms of eukaryotic RNA polymerases and which has been sequenced in budding yeast (gene RPB6 or RPO26), in fission yeast (gene rpb6 or rpo15), in human and in African swine fever virus [1] is evolutionary related [2] to archaeobacterial subunit K (gene rpoK). The archaeobacterial protein is colinear with the C-terminal part of the eukaryotic subunit.

Consensus pattern: [ST]-x-[FY]-E-x-[AT]-R-x-[LIVM]-[GSA]-x-R-[SA]-x-Q

- [1] Lu Z., Kutish G.F., Sussman M.D., Rock D.L. Nucleic Acids Res. 21:2940-2940(1993).
 [2] McKune K., Woychik N.A. J. Bacteriol. 176:4754-4756(1994).

5

483. RNA polymerases L / 13 to 16 Kd subunits signature

In eukaryotes, there are three different forms of DNA-dependent RNA polymerases (EC 2.7.7.6) transcribing different sets of genes. Each class of RNA polymerase is an assemblage of ten to twelve different polypeptides. In archaebacteria, there is generally a single form of RNA polymerase which also consist of an oligomeric assemblage of 10 to 13 polypeptides. It has been shown that small subunits of about 13 to 16 Kd found in all three types of eukaryotic polymerases are highly conserved. Subunits known to belong to this family are: - Budding yeast RPC19 subunit from RNA polymerases I and III [1]. - Budding yeast RPB11 subunit from RNA polymerase II [2]. - Mammalian RPB11 (gene POLR2K) from RNA polymerase II. - Caenorhabditis elegans hypothetical protein F58A4.9. - Methanococcus jannaschii RNA polymerase subunit L (gene rpoL). - Sulfolobus acidocaldarius RNA polymerase subunit L (gene rpoL) [3]. As a signature pattern a conserved region was selected which is located at the N-terminal extremity of these polymerase subunits; this region contains two cysteines that could play a role in the binding of a metal ion.

Consensus pattern: [DE](2)-H-[ST]-[LIVM]-[GAP]-N-x(11)-V-x-[FM]-x(2)-Y-x(3)-H-P

- [1] Dequard-Chablat M., Riva M., Carles C., Sentenac A. J. Biol. Chem. 266:15300-15307(1991).
 [2] Woychik N.A., McKune K., Lane W.S., Young R.A. Gene Expr. 3:77-82(1993).
 [3] Langer D. EMBL/GenBank: X70805.

484. RNA polymerases N / 8 Kd subunits signature

In eukaryotes, there are three different forms of DNA-dependent RNA polymerases (EC 2.7.7.6) transcribing different sets of genes. Each class of RNA polymerase is an assemblage of ten to twelve different polypeptides. In archaebacteria, there is generally a single form of RNA polymerase which also consist of an oligomeric assemblage of 10 to 13 polypeptides.

444

Archaeobacterial subunit N (gene rpoN) [1] is a small protein of about 8 Kd, it is evolutionary related [2] to a 8.3 Kd component shared by all three forms of eukaryotic RNA polymerases (gene RPB10 in yeast and POLR2J in mammals) as well as to African swine fever virus protein CP80R [3]. As a signature pattern a conserved region was selected which is located at the N-terminal extremity of these polymerase subunits; this region contains two cysteines that could play a role in the binding of a metal ion.

Consensus pattern: [LIVMF](2)-P-[LIVM]-x-C-F-[ST]-C-G-

[1] Langer D., Hain J., Thuriaux P., Zillig W. Proc. Natl. Acad. Sci. U.S.A. 92:5768-5772(1995).

[2] McKune K., Woychik N.A. J. Bacteriol. 176:4754-4756(1994).

[3] Yanez R.J., Rodriguez J.M., Nogal M.L., Yuste L., Enriquez C., Rodriguez J.F., Vinuela E. Virology 208:249-278(1995).

485. Ribonuclease HII

[1] Mian IS; Nucleic Acids Res 1997;25:3187-3189.

486. Ribonuclease PH signature

Prokaryotic ribonuclease PH (EC 2.7.7.56) (RNase PH) [1] is a phosphorolytic exoribonuclease that removes nucleotide residues following the -CCA terminus of tRNA and adds nucleotides to the ends of RNA molecules by using nucleoside diphosphates as substrates. RNase PH is a conserved protein of about 240 amino-acid residues. It is evolutionary related to *Caenorhabditis elegans* hypothetical protein B0564.1. As a signature pattern, the most highly conserved region was selected which is located in the central part of these proteins.

Consensus sequence: C-[DE]-[LIVM](2)-Q-[GTA]-D-G-[SG]-x(2)-[TA]-A

[1] Kelly K.O., Deutscher M.P. J. Biol. Chem. 267:17153-17158(1992).

487. RanBP1 domain

[1] Di Matteo G, Fuschi P, Zerfass K, Moretti S, Ricordy R, Cenciarelli C, Tripodi M, Jansen-Durr P, Lavia P; Cell Growth Differ 1995;6:1213-1224.

5 488. Rhodanese signatures

Rhodanese (thiosulfate sulfurtransferase) (EC 2.8.1.1) [1,2] is an enzyme which catalyzes the transfer of the sulfane atom of thiosulfate to cyanide, to form sulfite and thiocyanate. In vertebrates, rhodanese is a mitochondrial enzyme of about 300 amino-acid residues involved in forming iron-sulfur complexes and cyanide detoxification. A cysteine residue takes part in the catalytic mechanism. Some bacterial proteins closely related to rhodanese are also thought to express a sulfotransferase activity. These are: - *Azotobacter vinelandii* rhdA. - *Escherichia coli* sseA [3]. - *Saccharopolyspora erythraea* cysA [4]. - *Synechococcus* strain PCC 7942 rhdA [5]. RhdA is a periplasmic protein probably involved in the transport of sulfur compounds. Two patterns for the rhodanese family were developed. They are based on highly conserved regions, one which is located in the N-terminal region, the other at the C-terminal extremity of the enzyme.

Consensus pattern: [FY]-x(3)-H-[LIV]-P-G-A-x(2)-[LIVF]

Consensus pattern: [FY]-[DEAP]-G-[SA]-W-x-E-[FYW]

[1] Westley J. Meth. Enzymol. 77:285-291(1981).

[2] Weiland K.L., Dooley T.P. Biochem. J. 275:227-231(1991).

[3] Rudd K.E. Unpublished observations (1993).

[4] Donadio S., Shafiee A., Hutchinson C.R. J. Bacteriol. 172:350-360(1990).

[5] Laudenbach D.E., Ehrhardt D., Green L., Grossman A.R. J. Bacteriol. 173:2751-2760(1991).

489. Ribonuclease III family signature

Prokaryotic ribonuclease III (EC 3.1.26.3) (gene rnc) [1] is an enzyme that digests double-stranded RNA. It is involved in the processing of ribosomal RNA precursors and of some mRNAs. RNase III is evolutionary related [2] to the following proteins: - Fission yeast pac1, a ribonuclease that probably inhibits mating and meiosis by degrading a specific mRNA

required for sexual development. - Yeast ribonuclease III (gene RNT1), a dsRNA-specific nuclease that cleaves eukaryotic preribosomal RNA at various sites. - *Caenorhabditis elegans* hypothetical protein F26E4.13. - *Paramecium bursaria* chloroella virus 1 protein A464R. - *Synechocystis* strain PCC 6803 hypothetical protein slr0346. - Fission yeast hypothetical protein SpAC8A4.08c, a protein with a N-terminal helicase domain and a C-terminal RNase III domain. - *Caenorhabditis elegans* hypothetical protein K12H4.8, a protein with the same structure as SpAC8A4.08c. These proteins share regions of sequence similarity; one of which is a highly conserved stretch of 9 residues which has been developed as a signature pattern.

Consensus pattern: [DEQ]-[RQ]-[LM]-E-[FYW]-[LV]-G-D-[SAR]-

[1] Nashimoto H., Uchida H. Mol. Gen. Genet. 201:25-29(1985).

[2] Mian I.S. Nucleic Acids Res. 25:3187-3195(1997).

490. Rieske iron-sulfur protein signatures

Ubiquinol-cytochrome c reductase (EC 1.10.2.2) (also known as the bc1 complex or complex III) is one of the electron transport chains of mitochondria and of some aerobic prokaryotes; it catalyzes the oxidoreduction of ubiquinol and cytochrome c. In the chloroplast of plants and in cyanobacteria plastoquinone-plastocyanin reductase (EC 1.10.99.1) (also known as the b6f complex) is functionally similar and catalyzes the oxidoreduction of plastoquinol and cytochrome f. One of the components of these electron transfer systems is an iron-sulfur protein with a 2Fe-2S cluster, which is called the Rieske protein [1,2]. The Rieske protein contains approximately 190 amino acid residues. The iron-sulfur cluster is complexed to the protein through cysteine and histidine residues. Two perfectly conserved regions in Rieske proteins contain all the residues that bind the iron-sulfur cluster. Both regions contain two cysteines and a histidine. The first cysteine and the histidine are 2Fe-2S ligands while the remaining cysteines form a disulfide bond [3]. Two conserved regions were selected as signature patterns.

Consensus pattern: C-[TK]-H-L-G-C-[LIVST] [The first C and the H are 2Fe-2S ligands]
[The second C is involved in a disulfide bond]

Consensus pattern: C-P-C-H-x-[GSA] [The first C and the H are 2Fe-2S ligands] [The second C is involved in a disulfide bond]

[1] Gatti F.L., Meinhardt S.W., Ohnishi T., Tzagoloff A. J. Mol. Biol. 205:421-435(1989).

5 [2] Kallas T., Spiller S., Malkin R. Proc. Natl. Acad. Sci. U.S.A. 85:5794-5798(1988).

[3] Iwata S., Saynovits M., Link T.A., Michel H. Structure 4:567-579(1996).

491. Ribosomal protein L1 signature

10 Ribosomal protein L1 is the largest protein from the large ribosomal subunit. In Escherichia coli, L1 is known to bind to the 23S rRNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1, 2], groups: - Eubacterial L1. - Algal and plant chloroplast L1. - Cyanelle L1. - Archaeobacterial L1. - Vertebrate L10A. - Yeast SSM1. As a signature pattern, the best conserved region was selected located in the central section of these proteins. It is located at the end of an alpha helix thought to be involved in RNA-binding.

Consensus pattern: [IM]-x(2)-[LIVA]-x(2,3)-[LIVM]-G-x(2)-[LMS]-[GSNH]-[PTKR]-[KRAV]-G-x-[LIMF]-P-[DENSTKQ]

20 [1] Nikonov S.V., Nevskaya N., Eliseikina I.A., Fomenkova N.P., Nikulin A., Ossina N., Garber M., Jonsson B.-H., Briand C., Al-Karadaghi S., Svensson L.A., Aevvarsson A., Liljas A. EMBO J. 15:1350-1359(1996).

[2] Olvera J., Wool I.G. 2.3.CO;2-"Biochem. Biophys. Res. Commun. 220:954-957(1996).

25

492. Ribosomal protein L10 signature

30 Ribosomal protein L10 is one of the proteins from the large ribosomal subunit. L10 is a protein of 162 to 185 amino-acid residues which has only been found so far in eubacteria. A conserved region located in the N-terminal section of these proteins was used as a signature pattern.

Consensus pattern: [DEH]-x(2)-[GS]-[LIVMF]-[STN]-[VA]-x-[DEQK]-[LIVMA]-x(2)-[LIM]-R

5 493. Ribosomal protein L10e signature

A number of eukaryotic and archaebacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of: - Vertebrate L10 (QM) [1]. - Plant L10. - Caenorhabditis elegans L10 (F10B5.1). - Yeast L10 (QSR1). - Methanococcus jannaschii MJ0543. These proteins have 174 to 232 amino-acid residues. A conserved region located in the central section was selected as a signature pattern.

Consensus pattern: R-x-A-[FYW]-G-K-[PA]-x-G-x(2)-A-R-V

[1] Chan Y.-L., Diaz J.-J., Denoroy L., Madjar J.-J., Wool I.G. 2.3.CO;2-"Biochem. Biophys. Res. Commun. 255:952-956(1996).

494. Ribosomal protein L11 signature

Ribosomal protein L11 is one of the proteins from the large ribosomal subunit. In Escherichia coli, L11 is known to bind directly to the 23S rRNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1,2], groups:

- Eubacterial L11.
- Plant chloroplast L11 (nuclear-encoded).
- Read algal chloroplast L11.
- Cyanelle L11.
- Archaeobacterial L11.
- Mammalian L12.
- Plants L12.
- Yeast L12 (YL15).

L11 is a protein of 140 to 165 amino-acid residues. A conserved region located in the C-terminal section of these proteins was selected as a signature pattern. In Escherichia coli, the

C-terminal half of L11 has been shown [3] to be in an extended and loosely folded conformation and is likely to be buried within the ribosomal structure.

Consensus pattern: [RKN]-x-[LIVM]-x-G-[ST]-x(2)-[SNQ]-[LIVM]-G-x(2)-[LIVM]-x(0,1)-
[DENG]

[1] Pucciarelli G., Remacha M., Ballesta J.P.G.; Nucleic Acids Res. 18:4409-4416(1990).

[2] Otake E., Hashimoto T., Mizuta K., Suzuki K.; Protein Seq. Data Anal. 5:301-313(1993).

[3] Choli T. Biochem. Int. 19:1323-1338(1989).

495. Ribosomal protein L7/L12 C-terminal domain

[1] Leijonmarck M, Liljas A; J Mol Biol 1987;195:555-579.

496. Ribosomal protein L13 signature

Ribosomal protein L13 is one of the proteins from the large ribosomal subunit.

In Escherichia coli, L13 is known to be one of the early assembly proteins of the 50S ribosomal subunit. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups: - Eubacterial L13.

- Plant chloroplast L13 (nuclear-encoded). - Red algal chloroplast L13.

- Archaeobacterial L13. - Mammalian L13a (Tum P198). - Yeast Rp22 and Rp23.

L11 is a protein of 140 to 250 amino-acid residues. As a signature pattern, a conserved region was selected located in the C-terminal section of these proteins.

Consensus pattern: [LIVM]-[KRV]-[GK]-M-[LIV]-[PS]-x(4,5)-[GS]-[NQEKRA]-x(5)-
[LIVM]-x-[AIV]-[LFY]-x-[GDN]

[1] Chan Y.-L., Olvera J., Glueck A., Wool I.G. J. Biol. Chem. 269:5589-5594(1994).

497. Ribosomal protein L13e signature

A number of eukaryotic ribosomal proteins can be grouped on the basis of sequence similarities [1]. One of these families consists of:

- Vertebrate L13 (was previously known as Breast Basic Conserved protein 1 (BBC1)).
- Drosophila L13.
- Plant L13.
- Yeast probable L13 (YM9375.11c).

These proteins have 199 to 218 amino-acid residues. As a signature pattern, a stretch of about 16 residues in the first third of these proteins selected.

-Consensus pattern: [KR]-Y-x(2)-K-[LIVM]-R-[STA]-G-[KR]-G-F-[ST]-L-x-E

[1] Olvera J., Wool I.G. Biochem. Biophys. Res. Commun. 201:102-107(1994).

498. Ribosomal protein L14 signature

Ribosomal protein L14 is one of the proteins from the large ribosomal subunit.

In eubacteria, L14 is known to bind directly to the 23S rRNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups:

- Eubacterial L14.
- Algal and plant chloroplast L14.
- Cyanelle L14.

- Archaeobacterial L14.
- Yeast L17A.
- Mammalian L23.
- Caenorhabditis elegans L23 (B0336.10).
- Higher eukaryotes mitochondrial L14.
- Yeast mitochondrial Yml38 (gene MRPL38).

L14 is a protein of 119 to 137 amino-acid residues. As a signature pattern, a conserved region located in the C-terminal half of these proteins was selected.

-Consensus pattern: [GA]-[LIV](3)-x(9,10)-[DNS]-G-x(4)-[FY]-x(2)-[NT]-x(2)-V-[LIV]

[1] Otake E., Hashimoto T., Mizuta K., Suzuki K. Protein Seq. Data Anal. 5:301-313(1993).

499. Ribosomal protein L15 signature

Ribosomal protein L15 is one of the proteins from the large ribosomal subunit.

In Escherichia coli, L15 is known to bind the 23S rRNA. It belongs to a family

451

of ribosomal proteins which, on the basis of sequence similarities [1],

groups: - Eubacterial L15. - Plant chloroplast L15 (nuclear-encoded).

- Archaeobacterial L15. - Vertebrate L27a. - Tetrahymena thermophila L29.

- Fungi L27a (L29, CRP-1, CYH2).

- 5 L15 is a protein of 144 to 154 amino-acid residues. As a signature pattern,
a conserved region was selected in the C-terminal section of these proteins.

-Consensus pattern: K-[LIVM](2)-[GASL]-x-[GT]-x-[LIVMA]-x(2,5)-[LIVM]-x-[LIVMF]-
x(3,4)-[LIVMFCA]-[ST]-x(2)-A-x(3)-[LIVM]-x(3)-G

10

[1] Otake E., Hashimoto T., Mizuta K., Suzuki K. Protein Seq. Data Anal. 5:301-
313(1993).

- 5 500. Ribosomal protein L15e signature

A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped
on the basis of sequence similarities [1]. One of these families consists of:

- Mammalian L15. - Insect L15. - Plant L15. - Yeast YL10 (L13) (Rp15r).
- Thermoplasma acidophilum L15.

20

These proteins have about 200 amino acid residues. As a signature pattern,
a conserved region was selected located in the central section.

-Consensus pattern: [DE]-[KR]-A-R-x-L-G-[FY]-x-[SAP]-x(2)-G-[LIVMFY](4)-R-x-R-
[IV]-x-R-G

[1] Zwickl P., Lupas A., Baumeister W.

- 25 Biochem. Biophys. Res. Commun. 209:684-688(1995).

501. Ribosomal protein L17 signature

Ribosomal protein L17 is one of the proteins from the large ribosomal subunit.

- 30 L17 belongs to a family of ribosomal proteins which, on the basis of sequence
similarities, groups: - Eubacterial L17.

- Yeast mitochondrial YmL8 (gene MRPL8).

Eubacterial L17 is a protein of 120 to 130 amino-acid residues. Yeast YmL8 is

452

twice larger (238 residues), the sequence of its N-terminal half is colinear with that of eubacterial L17. As a signature pattern, a conserved region in the N-terminal section was selected.

-Consensus pattern: I-x-[ST]-[GT]-x(2)-[KR]-x-K-x(6)-[DE]-x-[LIMV]-[LIVMT]-T-x-[STAG]-[KR]

502. Ribosomal protein L18e signature

A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:

- Vertebrate L18 (known as L14 in *Xenopus*) [1]. - Plant L18.
- Yeast L18 (Rp28). - *Halobacterium marismortui* H129.
- *Sulfolobus acidocaldarius* H129e.

These proteins have 115 to 187 amino-acid residues., A stretch of about 13 residues in the first third of these proteins has been selected as a signature pattern.

-Consensus pattern: [KRE]-x-L-x(2)-[PS]-[KR]-x(2)-[RH]-[PSA]-x-[LIVM]-[NS]-[LIVM]-x-[RK]-[LIVM]

[1] Puder M., Barnard G.F., Staniunas R.J., Steele G.D. Jr., Chen L.B. *Biochim. Biophys. Acta* 1216:134-136(1993).

503. Ribosomal L18p family

It has been shown that the amino terminal 93 amino acids of Swiss:P09895 are necessary and sufficient to bind 5S rRNA in vitro. The carboxyl-terminal half of the protein, comprising amino acids 151-296, serves to localize the protein to the nucleolus [1].

Number of members: 26

[1]

Medline: 96212235

Distinct domains in ribosomal protein L5 mediate 5 S rRNA binding and nucleolar localization.

Michael WM, Dreyfuss G;

J Biol Chem 1996;271:11571-11574.

504. Ribosomal protein L19 signature

5 Ribosomal protein L19 is one of the proteins from the large ribosomal subunit.

In *Escherichia coli*, L19 is known to be located at the 30S-50S ribosomal subunit interface and may play a role in the structure and function of the aminoacyl-tRNA binding site. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities, groups: - Eubacterial L19.

10 - Red algal chloroplast L19. - Cyanelle L19.

L19 is a protein of 120 to 130 amino-acid residues.,

A conserved region in the C-terminal section has been selected as a signature pattern.

-Consensus pattern: [LIVM]-x-[KRGTI]-x-[GSAI]-[KRQDA]-[VG]-[RSN]-X(0,1)-[KR]-
[SA]-[KY]-[KLI]-[LYS]-Y-[LIM]-R

505. Ribosomal protein L19e signature

A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:

- Mammalian ribosomal protein L19 [1]. - *Drosophila* ribosomal protein L19 [2].
- Slime mold (*D. discoideum*) vegetative specific protein V14 [3].
- Yeast ribosomal protein L19 (YL14). - Archeobacterial ribosomal protein L19E.

These proteins have 148 to 203 amino-acid residues.

A stretch of about 20 residues in the N-terminal part of these

25 proteins has been selected as a signature pattern.

-Consensus pattern: Q-[KR]-R-[LIVM]-x-[SA]-x(4)-[CV]-G-x(3)-[IV]-[WK]-[LIVF]-
[DN]-P

[1] Chan Y.-L., Lin A., McNally J., Peleg D., Meyuhas O., Wool I.G.

J. Biol. Chem. 262:1111-1115(1987).[2] Hart K., Klein T., Wilcox M.

30 Mech. Dev. 43:101-110(1993).[3] Singleton C.K., Manning S.S., Ken R.

Nucleic Acids Res. 17:9679-9692(1989).

506. Ribosomal protein L1e signature (Ribosomal_L4)

A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists [1,2,3,4] of: - Vertebrate L1 (L4). - Drosophila L1. - Plant L1. - Yeast L2 (Rp2).

- Fission yeast L2. - Halobacterium marismortui HmaL4 (HL6).
- Methanococcus jannaschii MJ0177.

These proteins have 246 (archaeobacteria) to 427 (human) amino acids. A conserved region in the N-terminal part of these proteins has been selected as a signature pattern.

-Consensus pattern: N-x(3)-[KRM]-x(2)-A-[LIVT]-x-S-A-[LIV]-x-A-[ST]-[SGA]-x(7)-[RK]-[GS]-H

[1] Rafti F., Gargiulo G., Manzi A., Malva C., Graziani F.

Nucleic Acids Res. 17:456-456(1989).[2] Presutti C., Villa T., Bozzoni I.

Nucleic Acids Res. 21:3900-3900(1993).

[3] Bagni C., Mariottini P., Annesi F., Amaldi F.

Biochim. Biophys. Acta 1216:475-478(1993).

[3] Arndt E., Kroemer W., Hatakeyama T. J. Biol. Chem. 265:3034-3039(1990).

507. Ribosomal protein L2 signature

Ribosomal protein L2 is one of the proteins from the large ribosomal subunit.

In Escherichia coli, L2 is known to bind to the 23S rRNA and to have peptidyltransferase activity. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1,2], groups: - Eubacterial L2.

- Algal and plant chloroplast L2. - Cyanelle L2. - Archaeobacterial L2.

- Plant L2. - Slime mold L2. - Marchantia polymorpha mitochondrial L2.

- Paramecium tetraurelia mitochondrial L2. - Fission yeast K5, K37 and KD4.

- Yeast YL6. - Vertebrate L8.

The best conserved region located in the C-terminal section of these proteins has been selected as

a signature pattern.

-Consensus pattern: P-x(2)-R-G-[STAIV](2)-x-N-[APK]-x-[DE]

[1] Marty I., Meyer Y.

Nucleic Acids Res. 20:1517-1522(1992).

[2] Otaka E., Hashimoto T., Mizuta K., Suzuki K.

Protein Seq. Data Anal. 5:301-313(1993).

5 508. Ribosomal protein L20 signature

Ribosomal protein L20 is one of the proteins from the large ribosomal subunit.

In Escherichia coli, L20 is known to bind directly to the 23S rRNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups: - Eubacterial L20. - Algal and plant chloroplast L20.

10 - Cyanelle L20.

L20 is a protein of about 120 amino-acid residues. A conserved region located in the central section of these proteins has been selected as a signature pattern.

-Consensus pattern: K-x(3)-[KRC]-x-[LIVM]-W-[IV]-[STNALV]-R-[LIVM]-[NS]-x(3)-[RKHS]

15 [1] Otaka E., Hashimoto T., Mizuta K., Suzuki K.

Protein Seq. Data Anal. 5:301-313(1993).

20 509. Ribosomal protein L21e signature

A number of eukaryotic and archaebacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:

- Mammalian L21 [1]. - Entamoeba histolytica L21 [2].

- Caenorhabditis elegans L21 (C14B9.7). - Yeast L21E (URP1) [3].

- Halobacterium marismortui HL31 [4].

25 These proteins have 160 (eukaryotes) or 95 (archebacteria) amino-acid residues. A conserved region in the central part of these proteins has been selected as a signature pattern.

-Consensus pattern: G-[DE]-x-V-x(10)-[GV]-x(2)-[FYH]-x(2)-[FY]-x-G-x-T-G

[1] Devi K.R.G., Chan Y.-L., Wool I.G.

30 Biochem. Biophys. Res. Commun. 162:364-370(1989).

[2] Petter R., Rozenblatt S., Nuchamowitz Y., Mirelman D.

Mol. Biochem. Parasitol. 56:329-333(1992).

[3] Jank B., Waldherr M., Schweyen R.J. Curr. Genet. 23:15-18(1993).

[4] Hatakeyama T., Kimura M. Eur. J. Biochem. 172:703-711(1988).

510. Ribosomal protein L21 signature

5 Ribosomal protein L21 is one of the proteins from the large ribosomal subunit.

In *Escherichia coli*, L21 is known to bind to the 23S rRNA in the presence of L20. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities, groups: - Eubacterial L21.

- *Marchantia polymorpha* chloroplast L21. - *Cyanelle* L21.

10 - *Spinach* chloroplast L21 (nuclear-encoded).

Eubacterial L21 is a protein of about 100 amino-acid residues, the mature form of the *spinach* chloroplast L21 has 200 residues. A conserved region located in the C-terminal section of these proteins has been selected as a signature pattern.

-Consensus pattern: [IVT]-x(3)-[KR]-x(3)-[KRQ]-K-x(6)-G-[HF]-R-[RQ]-x(2)-[ST]

511. Ribosomal protein L22 signature

Ribosomal protein L22 is one of the proteins from the large ribosomal subunit.

In *Escherichia coli*, L22 is known to bind 23S rRNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1,2,3], groups: - Eubacterial L22.

- Algal and plant chloroplast L22 (in legumes L22 is encoded in the nucleus instead of the chloroplast). - *Cyanelle* L22. - Archaeobacterial L22.

- Mammalian L17. - Plant L17. - Yeast YL17.

25 A conserved region located in the C- terminal section of these proteins has been selected as a signature pattern.

-Consensus pattern: [RKQN]-x(4)-[RH]-[GAS]-x-G-[KRQS]-x(9)-[HDN]-[LIVM]-x-[LIVMS]-x-[LIVM]

[1] Gantt J.S., Baldauf S.L., Calie P.J., Weeden N.F., Palmer J.D.

30 EMBO J. 10:3073-3078(1991).[2] Madsen L.H., Kreiberg J.D., Gausing K. Curr. Genet. 19:417-422(1991).

[3] Otaka E., Hashimoto T., Mizuta K., Suzuki K. Protein Seq. Data Anal. 5:301-313(1993).

512. Ribosomal protein L23 signature

Ribosomal protein L23 is one of the proteins from the large ribosomal subunit.

In *Escherichia coli*, L23 is known to bind a specific region on the 23S rRNA; in yeast, the corresponding protein binds to a homologous site on the 26S rRNA [1]. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [2,3,4], groups: - Eubacterial L23.

- Algal and plant chloroplast L23. - Archaeobacterial L23. - Mammalian L23A.

- *Caenorhabditis elegans* L23A (F55D10.2). - Fungi L25.

- Yeast mitochondrial YmL41 (gene MRPL41 or MRP20).

A small conserved region in the C-terminal section of these proteins, which is probably involved in rRNA-binding has been selected as a signature pattern [2].

-Consensus pattern: [RK](2)-[AM]-[IVFYT]-[IV]-[RKT]-L-[STANEQK]-x(7)-[LIVMFT]

[1] El Baradi T.T.A.L., Raue H.A., van de Regt C.H.F., Verbree E.C.,

Planta R.J. EMBO J. 4:210-2107(1985).

[2] Raue H.A., Otaka E., Suzuki K. J. Mol. Evol. 28:418-426(1989).

[3] Fearon K., Mason T.L. J. Biol. Chem. 267:5162-5170(1992).

[4] Otaka E., Hashimoto T., Mizuta K.

Protein Seq. Data Anal. 5:285-300(1993).

513. Ribosomal protein L24 signature

Ribosomal protein L24 is one of the proteins from the large ribosomal subunit.

L24 belongs to a family of ribosomal proteins which, on the basis of sequence similarities, groups: - Eubacterial L24.

- Plant chloroplast L24 (nuclear-encoded). - Red algal L24. - Vertebrate L26.

- Yeast L26 (YL33). - Archaeobacterial HmaL24 (HL15).

- A probable ribosomal protein from *Sulfolobus acidocaldarius* [1].

In their mature form, these proteins have 103 to 150 amino-acid residues.

A conserved stretch of 20 residues in their N-terminal section has been selected as a signature pattern.

458

-Consensus pattern: [GDEN]-D-x-V-x-[IV]-[LIVMA]-x-G-x(2)-[KRA]-[GNQ]-x(2,3)-
[GA]-x-[IV]

[1] Ouzounis C., Kyripides N., Sander C.

Nucleic Acids Res. 23:565-570(1995).

5

514. Ribosomal protein L24e signature

A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists [1] of:

10

- Mammalian ribosomal protein L24.
- Yeast ribosomal protein L30A/B (Rp29) (YL21).
- Kluyveromyces lactis ribosomal protein L30.
- Arabidopsis thaliana ribosomal protein L24 homolog.
- Haloarcula marismortui ribosomal protein HL21/HL22.
- Methanococcus jannaschii MJ1201.

15

These proteins have 60 to 160 amino-acid residues. The most conserved region, which is located in the N-terminal region of these proteins has been selected as a signature pattern.

-Consensus pattern: [FY]-x-[GSH]-x(2)-[IV]-x-P-G-x-G-x(2)-[FYV]-x-[KRHE]-x-D

[1] Chan Y.-L., Olvera J., Wool I.G.

20

Biochem. Biophys. Res. Commun. 202:1176-1180(1994).

515. Ribosomal protein L27 signature

Ribosomal protein L27 is one of the proteins from the large ribosomal subunit.

25

L27 belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1,2], groups: - Eubacterial L27.

- Plant chloroplast L27 (nuclear-encoded). - Algal chloroplast L27.
- Yeast mitochondrial YmL2 (gene MRPL2 or MRP7).

The schematic relationship between these groups of proteins is shown below.

30

Eub. L27 NxxxxxxxxxAlgal L27 Nxxxxxxxxx

Plant L27 ttttNxxxxxxxxxxxxxxxx

Yeast MRP7 tttNxx

***'t': transit peptide.

'N': N-terminal of mature protein. '*': position of the pattern.

-Consensus pattern: G-x-[LIVM](2)-x-R-Q-R-G-x(5)-G

[1] Elhag G.A., Bourque D.P. Biochemistry 31:6856-6864(1992).

[2] Otake E., Hashimoto T., Mizuta K.

5 Protein Seq. Data Anal. 5:285-300(1993).

516. Ribosomal L28 family

The ribosomal 28 family includes L28 proteins from bacteria
and chloroplasts. The L24 protein from yeast Swiss:P36525
also contains a region of similarity to prokaryotic L28
proteins. L24 from yeast is also found in the large
ribosomal subunit

Number of members: 24

517. Ribosomal protein L29 signature

Ribosomal protein L29 is one of the proteins from the large ribosomal subunit.
L29 belongs to a family of ribosomal proteins which, on the basis of sequence
similarities [1], groups: - Eubacterial L29. - Red algal L29.

- Archaeobacterial L29. - Mammalian L35 - Caenorhabditis elegans L35 (ZK652.4).

- Yeast L35.

L29 is a protein of 63 to 138 amino-acid residues.

A conserved region located in the central section of L29 has been selected as a
signature pattern.

-Consensus pattern: [KNQS]-[PSTL]-x(2)-[LIMFA]-[KRGSA]-x-[LIVYSTA]-[KR]-
[KRHQS]-[DESTANRL]-[LIV]-A-[KRCQVT]-[LIVMA]

[1] Otake E., Hashimoto T., Mizuta K.

Protein Seq. Data Anal. 5:285-300(1993).

518. Ribosomal protein L3 signature

Ribosomal protein L3 is one of the proteins from the large ribosomal subunit.

In *Escherichia coli*, L3 is known to bind to the 23S rRNA and may participate in the formation of the peptidyltransferase center of the ribosome. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1,2,3,4], groups: - Eubacterial L3. - Red algal L3. - Cyanelle L3.

- 5 - Archaeobacterial *Halobacterium marismortui* HmaL3 (HL1).
- Yeast L3 (also known as trichodermin resistance protein) (gene TCM1).
- *Arabidopsis thaliana* L3 (genes ARP1 and ARP2). - Mammalian L3 (L4).
- Mammalian mitochondrial L3. - Yeast mitochondrial YmL9 (gene MRPL9).

A conserved region located in the central section of these proteins has been selected
10 as a signature pattern.

-Consensus pattern: [FL]-x(6)-[DN]-x(2)-[AGS]-x-[ST]-x-G-[KRH]-G-x(2)-G-x(3)-R

[1] Arndt E., Kroemer W., Hatakeyama T. J. Biol. Chem. 265:3034-3039(1990).

[2] Graack H.-R., Grohmann L., Kitakawa M., Schaefer K.L., Kruft V.

Eur. J. Biochem. 206:373-380(1992).

15 [3] Herwig S., Kruft V., Wittmann-Liebold B.

Eur. J. Biochem. 207:877-885(1992).

[4] Otake E., Hashimoto T., Mizuta K., Suzuki K.

Protein Seq. Data Anal. 5:301-313(1993).

20 519. Ribosomal protein L30 signature

Ribosomal protein L30 is one of the proteins from the large ribosomal subunit. L30 belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups: - Eubacterial L30. - Archaeobacterial L30.

- 25 - *Drosophila* L7. - Slime mold L7. - Mammalian L7. - Fungi L7 (YL8).
- Yeast mitochondrial L33.

L30 from eubacteria are small proteins of about 60 residues, those from archaeobacteria are proteins of about 150 residues. Eukaryotic L7 are proteins of about 250 to 270 residues. The schematic relationship between the three
30 groups of proteins is shown below. Eub. L30 NxxxxxxxxxxC

Arc. L30 NxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxC

Euk. L7 NxxC

*****!': position of the pattern.

461

The signature pattern for this family of ribosomal proteins spans the N-terminal half of the region common to all these proteins.

-Consensus pattern: [IVT]-[LIVM]-x(2)-[LF]-x-[LI]-x-[KRHQEG]-x(2)-[STNQH]-x-[IVT]-x(10)-[LMS]-[LIV]-x(2)-[LIVA]-x(2)-[LMFY]-[IVT]

- 5 [1] Mizuta K., Hashimoto T., Otake E.
Nucleic Acids Res. 20:1011-1016(1992).

520. Ribosomal protein L31 signature

- 10 Ribosomal protein L31 is one of the proteins from the large ribosomal subunit. L31 is a protein of 66 to 97 amino-acid residues which has only been found so far in eubacteria and in some algal chloroplasts.
A conserved region located in the central section of these proteins has been selected as a signature pattern.

15 -Consensus pattern: H-P-F-[FY]-[TI]-x(9)-G-R-[AIV]-x-[KRQ]

521. Ribosomal protein L31e signature

- 20 A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:

- Mammalian L31 [1]. - Chlamydomonas reinhardtii L31. - Yeast L34.
- Halobacterium marismortui HL30 [2].

These proteins have 87 to 128 amino-acid residues.

A conserved region, located in the central section has been selected as a signature pattern.

25 -Consensus pattern: V-[KR]-[LIVM]-x(3)-[LIVM]-N-x-[AKH]-x-W-x-[KR]-G

- [1] Tanaka T., Kuwano Y., Kuzumaki T., Ishikawa K., Ogata K.
Eur. J. Biochem. 162:45-48(1987).[2] Bergmann U., Arndt E.
Biochim. Biophys. Acta 1050:56-60(1990).

30

522. Ribosomal protein L33 signature

Ribosomal protein L33 is one of the proteins from the large ribosomal subunit. In Escherichia coli, L33 has been shown to be on the surface of 50S subunit.

462

L33 belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1,2,3], groups: - Eubacterial L33.

- Algal and plant chloroplast L33. - Cyanelle L33.

L33 is a small protein of 49 to 66 amino-acid residues. A conserved region located in the central section of L33 has been selected as a signature pattern.

-Consensus pattern: Y-x-[ST]-x-[KR]-[NS]-x(4)-[PATQ]-x(1,2)-[LIVM]-[EA]-x(2)-K-[FY]-[CSD]

[1] Kruft V., Kapp U., Wittmann-Liebold B. Biochimie 73:855-860(1991).

[2] Sharp P.M. Gene 139:129-130(1994).

[3] Otaka E., Hashimoto T., Mizuta K.

Protein Seq. Data Anal. 5:285-300(1993).

523. Ribosomal protein L34 signature

Ribosomal protein L34 is one of the proteins from the large subunit of the prokaryotic ribosome. It is a small basic protein of 44 to 51 amino-acid residues [1]. L34 belongs to a family of ribosomal proteins which, on the basis of sequence similarities, groups: - Eubacterial L34.

- Red algal chloroplast L34. - Cyanelle L34.

A conserved region that corresponds to the N-terminal half of L34 has been selected as a signature pattern.

-Consensus pattern: K-[RG]-T-[FYWL]-[EQS]-x(5)-[KRHS]-x(4,5)-G-F-x(2)-R

[1] Old I.G., Margarita D., Saint Girons I.

Nucleic Acids Res. 20:6097-6097(1992).

524. Ribosomal protein L34e signature

A number of eukaryotic and archaebacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:

- Mammalian L34. - Mosquito L31 [1]. - Plant L34 [2].

- Yeast putative ribosomal protein YIL052c. - Methanococcus jannaschii MJ0655.

These proteins have 89 to 129 amino-acid residues.

A conserved region located in the N-terminal section of these proteins has been selected as a signature pattern.

463

-Consensus pattern: Y-x-[ST]-x-S-[NY]-x(5)-[KR]-T-P-G

[1] Lan Q., Niu L.L., Fallon A.M.

Biochim. Biophys. Acta 1218:460-462(1994).

[2] Gao J., Kim S.R., Chung Y.Y., Lee J.M., An G.

5 Plant Mol. Biol. 25:761-770(1994).

525. Ribosomal protein L35Ae signature

A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped
10 on the basis of sequence similarities. One of these families consists of:

- Vertebrate L35A. - *Caenorhabditis elegans* L35A (F10E7.7).

- Yeast L37A/L37B (Rp47). - *Pyrococcus woesei* L35A homolog [1].

These proteins have 87 to 110 amino-acid residues.

A highly conserved stretch of 22 residues in the C-terminal part of
15 these proteins has been selected as a signature pattern.

-Consensus pattern: G-K-[LIVM]-x-R-x-H-G-x(2)-G-x-V-x-A-x-F-x(3)-[LI]-P

[1] Ouzounis C., Kyripides N., Sander C.

Nucleic Acids Res. 23:565-570(1995).

20 526. Ribosomal protein L36 signature

Ribosomal protein L36 is the smallest protein from the large subunit of the prokaryotic
ribosome. It belongs to a family of ribosomal proteins which, on the basis of sequence
similarities [1], groups: - Eubacterial L36. - Algal and plant chloroplast L36. - Cyanelle

25 L36. L36 is a small basic and cysteine-rich protein of 37 amino-acid residues. As a signature
pattern, a conserved region that corresponds to positions 11 to 36 in L36 and includes three
conserved cysteine residues has been developed.

Consensus pattern: C-x(2)-C-x(2)-[LIVM]-x-R-x(3)-[LIVMN]-x-[LIVM]-x-C-x(3,4)- [KR]-
H-x-Q-x-Q-

30 [1] Otaka E., Hashimoto T., Mizuta K. Protein Seq. Data Anal. 5:285-300(1993).

527. Ribosomal protein L36e signature

A number of eukaryotic ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of: - Mammalian L36 [1].

- Drosophila L36 (M(1)1B). - Caenorhabditis elegans L36 (F37C12.4).

- Candida albicans L39. - Yeast YL39.

5 These proteins have 99 to 104 amino acids.

A conserved region in the central part of these proteins has been selected as a signature pattern.

-Consensus pattern: P-Y-E-[KR]-R-x-[LIVM]-[DE]-[LIVM](2)-[KR]

[1] Chan Y.-L., Paz V., Olvera J., Wool I.G.

10 Biochem. Biophys. Res. Commun. 192:849-853(1993).

528. Ribosomal protein L39e signature

A number of eukaryotic and archaebacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:

- Mammalian L39 [1]. - Plants L39. - Yeast L46 [2]. - Archebacterial L39e [3].

These proteins are very basic. About 50 residues long, they are the smallest proteins of eukaryotic-type ribosomes. A conserved region in the C-terminal section of these proteins has been selected as a signature pattern.

-Consensus pattern: [KRA]-T-x(3)-[LIVM]-[KRQF]-x-[NHS]-x(3)-R-[NHY]-W-R-R

[1] Lin A., McNally J., Wool I.G. J. Biol. Chem. 259:487-490(1984).

[2] Leer R.J., van Raamsdonk-Duin M.M.C., Kraakman P., Mager W.H., Planta R.J. Nucleic Acids Res. 13:701-709(1985).

[3] Ramirez C., Louie K.A., Matheson A.T. FEBS Lett. 250:416-418(1989).

529. Ribosomal L40e family

Bovine L40 has been identified as a secondary RNA binding protein [1]. L40 is fused to a ubiquitin protein [2].

30 Number of members: 27

[1]

Medline: 88203200

RNA binding proteins of the large subunit of bovine

mitochondrial ribosomes.

Piatyszek MA, Denslow ND, O'Brien TW;

Nucleic Acids Res 1988;16:2565-2583.

[2]Medline: 96011832

- 5 The carboxyl extensions of two rat ubiquitin fusion proteins
are ribosomal proteins S27a and L40.

Chan YL, Suzuki K, Wool IG;

Biochem Biophys Res Commun 1995;215:682-690.

10

530. (Ribosomal L44) Ribosomal protein L44e signature

A number of eukaryotic and archaebacterial ribosomal proteins can be grouped
on the basis of sequence similarities. One of these families consists of:

- Mammalian L44 [1]. - Trypanosoma brucei L44.
- Caenorhabditis elegans L44 (C09H10.2). - Fungal L44 (L41).
- Halobacterium marismortui LA [2].

These proteins have 92 to 105 amino-acid residues.

A conserved region located in the C-terminal part of these proteins has been
selected as a signature pattern.

-Consensus pattern: K-x-[TV]-K-K-x(2)-L-[KR]-x(2)-C

[1] Gallagher M.J., Chan Y.-L., Lin A., Wool I.G. DNA 7:269-273(1988).

[2] Bergmann U., Wittmann-Liebold B.

Biochim. Biophys. Acta 1173:195-200(1993)

25

531. Ribosomal protein L5 signature

Ribosomal protein L5 is one of the proteins from the large ribosomal subunit.

In Escherichia coli, L5 is known to be involved in binding 5S RNA to the large
ribosomal subunit. It belongs to a family of ribosomal proteins which, on the

30 basis of sequence similarities [1,2,3,4], groups: - Eubacterial L5.

- Algal chloroplast L5. - Cyanelle L5. - Archaebacterial L5. - Mammalian L11.
- Tetrahymena thermophila L21. - Slime mold L5 (V18). - Yeast L16 (39A).
- Plants mitochondrial L5.

466

L5 is a protein of about 180 amino-acid residues.

A conserved region, located in the first third of these proteins has been selected as a signature pattern.

-Consensus pattern: [LIVM]-x(2)-[LIVM]-[STAVC]-[GE]-[QV]-x(2)-[LIVMA]-x-[STC]-x-[STAG]-[KRH]-x-[STA]

[1] Hatakeyama T., Hatakeyama T. Biochim. Biophys. Acta 1039:343-347(1990).

[2] Rosendahl G., Andreasen P.H., Kristiansen K. Gene 98:161-167(1991).

[3] Yang D., Gunther I., Matheson A.T., Auer J., Spicker G., Boeck A. Biochimie 73:679-682(1991).

[4] Otaka E., Hashimoto T., Mizuta K., Suzuki K. Protein Seq. Data Anal. 5:301-313(1993).

532. ribosomal L5P family C-terminus

This region is found associated with Ribosomal_L5.

Number of members: 60

533. Ribosomal protein L6 signatures

Ribosomal protein L6 is one of the proteins from the large ribosomal subunit. In Escherichia coli, L6 is known to bind directly to the 23S rRNA and is located at the aminoacyl-tRNA binding site of the peptidyltransferase center. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1,2,3,4], groups: - Eubacterial L6.

- Algal chloroplast L6.
- Cyanelle L6.
- Archaeobacterial L6.
- Marchantia polymorpha mitochondrial L6.
- Yeast mitochondrial YmL6 (gene MRPL6).
- Mammalian L9.
- Drosophila L9.
- Plants L9.
- Yeast L9 (YL11).

While all the above proteins are evolutionary related it is very difficult to derive a pattern that will find them all. Two patterns were therefore created, the first to detect eubacterial, cyanobacterial and mitochondrial L6, the second to detect archaeobacterial L6 as well as eukaryotic L9.

- 5 -Consensus pattern: [PS]-[DENS]-x-Y-K-[GA]-K-G-[LIVM]
 -Consensus pattern: Q-x(3)-[LIVM]-x(2)-[KR]-x(2)-R-x-F-x-D-G-[LIVM]-Y-[LIVM]-x(2)-[KR]

[1] Suzuki K., Olvera J., Wool I.G. Gene 93:297-300(1990).

10 [2] Schwank S., Harrer R., Schueller H.-J., Schweizer E. Curr. Genet. 24:136-140(1993).

[3] Golden B.L., Ramakrishnan V., White S.W. EMBO J. 12:4901-4908(1993).

[4] Otaka E., Hashimoto T., Mizuta K., Suzuki K. Protein Seq. Data Anal. 5:301-313(1993).

15 534. Ribosomal protein L6e signature

A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:

- Mammalian ribosomal protein L6 (L6 was previously known as TAX-responsive enhancer element binding protein 107).
- 20 - Caenorhabditis elegans ribosomal protein L6 (R151.3).
- Yeast ribosomal protein YL16A/YL16B.
- Mesembryanthemum crystallinum ribosomal protein YL16-like.

These proteins have 175 (yeast) to 287 (mammalian) amino acids. A highly conserved region in the central part of these proteins has been selected as a signature
 25 pattern.

-Consensus pattern: N-x(2)-P-L-R-R-x(4)-[FY]-V-I-A-T-S-x-K

535. Ribosomal protein L7Ae signature

30 A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:

- Vertebrate L7A (SURF3) [1]. - Plant L7A. - Yeast L7A (YL5) (Rp6).
- Yeast protein NHP2 [2]. - Yeast hypothetical protein YEL026w.

- *Bacillus subtilis* hypothetical protein ylxQ. - *Halobacterium marismortui* Hs6.
- *Methanococcus jannaschii* MJ1203.

These proteins have 100 to 265 amino-acid residues.

A conserved region located in the central section has been selected as a signature pattern.

5 -Consensus pattern: [CA]-x(4)-[IV]-P-[FY]-x(2)-[LIVM]-x-[GSQ]-[KRQ]-x(2)-L-G

[1] Colombo P., Yon J., Garson K., Fried M.

Proc. Natl. Acad. Sci. U.S.A. 89:6358-6362(1992).

[2] Kolodrubetz D., Burgum A. Yeast 7:79-90(1991).

10

536. Ribosomal protein L9 signature

Ribosomal protein L9 is one of the proteins from the large ribosomal subunit.

In *Escherichia coli*, L9 is known to bind directly to the 23S rRNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities

15 [1,2], groups: - Eubacterial L9. - Cyanobacterial L9.

- Plant chloroplast L9 (nuclear-encoded). - Red algal chloroplast L9.

A conserved region, located in the N-terminal section of these proteins has been selected as a signature pattern.

-Consensus pattern: G-x(2)-[GN]-x(4)-V-x(2)-G-[FY]-x(2)-N-[FY]-L-x(5)-[GA]-
20 x(3)-[STN]

[1] Hoffman D.W., Davies C., Gerchman S.E., Kycia J.H., Porter S.J.,

White S.W., Ramakrishnan V. EMBO J. 13:205-212(1994).

[2] Otake E., Hashimoto T., Mizuta K., Suzuki K.

Protein Seq. Data Anal. 5:301-313(1993).

25

537. Ribosomal protein S10 signature

Ribosomal protein S10 is one of the proteins from the small ribosomal subunit.

In *Escherichia coli*, S10 is known to be involved in binding tRNA to the
30 ribosomes. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups: - Eubacterial S10.

- Algal chloroplast S10. - *Cyanella* S10. - Archaeobacterial S10.

- *Marchantia polymorpha* and *Prototheca wickerhamii* mitochondrial S10.

469

- Arabidopsis thaliana mitochondrial S10 (nuclear encoded). - Vertebrate S20.
- Plant S20. - Yeast URP2.

S10 is a protein of about 100 amino-acid residues.

A conserved region located in the center of these proteins has been selected as a signature pattern.

-Consensus pattern: [AV]-x(3)-[GDNSR]-[LIVMSTA]-x(3)-G-P-[LIVM]-x-[LIVM]-P-T
 [1] Otake E., Hashimoto T., Mizuta K.
 Protein Seq. Data Anal. 5:285-300(1993).

538. Ribosomal protein S11 signature

Ribosomal protein S11 [1] plays an essential role in selecting the correct tRNA in protein biosynthesis. It is located on the large lobe of the small ribosomal subunit. S11 belongs to a family of ribosomal proteins which, on the basis of sequence similarities, groups [2]: - Eubacterial S11.

- Algal and plant chloroplast S11. - Cyanelle S11. - Archaeobacterial S11.
- Marchantia polymorpha and Prototheca wickerhamii mitochondrial S11.
- Acanthamoeba castellanii mitochondrial S11. - Neurospora crassa S14 (crp-2).
- Yeast S14 (RP59 or CRY1).
- Mammalian, Drosophila, Trypanosoma, and plant S14.
- Caenorhabditis elegans S14 (F37C12.9).

One of the best conserved regions in these proteins was selected as a signature pattern.

-Consensus pattern: [LIVMF]-x-[GSTAC]-[LIVMF]-x(2)-[GSTAL]-x(0,1)-[GSN]-
 [LIVMF]-x-[LIVM]-x(4)-[DEN]-x-T-P-x-[PA]-[STCH]-[DN]
 [1] Kimura M., Kimura J., Hatakeyama T. FEBS Lett. 240:15-20(1988).
 [2] Otake E., Hashimoto T., Mizuta K.
 Protein Seq. Data Anal. 5:285-300(1993).

539. Ribosomal protein S12 signature

Ribosomal protein S12 is one of the proteins from the small ribosomal subunit.

In Escherichia coli, S12 is known to be involved in the translation initiation

470

step. It is a very basic protein of 120 to 150 amino-acid residues. S12 belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups: - Eubacterial S12. - Archaeobacterial S12.

- Algal and plant chloroplast S12. - Cyanelle S12.

5 - Protozoa and plant mitochondrial S12. - Yeast S28.

- Drosophila mitochondrial protein tko (Technical KnockOut). - Mammalian S23.

The best conserved regions in these proteins, located in the center of each sequence have been selected as a signature pattern.

-Consensus pattern: [RK]-x-P-N-S-[AR]-x-R

10 [1] Otake E., Hashimoto T., Mizuta K.

Protein Seq. Data Anal. 5:285-300(1993).

540. Ribosomal protein S12e signature

15 A number of eukaryotic ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of: - Vertebrate S12 [1].

- Trypanosoma brucei S12 [2]. - Caenorhabditis elegans S12 (F54E7.2).

- Drosophila S12. - Yeast S12.

These proteins have 130 to 150 amino acids.

20 A conserved region in the N-terminal part of these proteins has been selected as a signature pattern.

-Consensus pattern: A-L-[KRQP]-x-V-L-x(2)-[SA]-x(3)-[DN]-G-L

[1] Lin A., Chan Y.-L., Jones R., Wool I.G.

J. Biol. Chem. 262:14343-14351(1987).[2] Marchal C., Ismaili N., Pays E.

25 Mol. Biochem. Parasitol. 57:331-334(1993).

541. Ribosomal protein S13 signature

Ribosomal protein S13 is one of the proteins from the small ribosomal subunit.

30 In Escherichia coli, S13 is known to be involved in binding fMet-tRNA and, hence, in the initiation of translation. It is a basic protein of 115 to 177 amino-acid residues and belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1,2], groups: - Eubacterial S13.

471

- Plant chloroplast S13 (nuclear encoded). - Red algal chloroplast S13.
- Cyanelle S13. - Archaeobacterial S13. - Plant mitochondrial S13.
- Mammalian and plant S18.

The best conserved regions in these proteins, located in their C-terminal
 5 part have been selected as a signature pattern.

-Consensus pattern: [KRQS]-G-x-R-H-x(2)-[GSNH]-x(2)-[LIVMC]-R-G-Q

[1] Chan Y.-L., Paz V., Wool I.G.

Biochem. Biophys. Res. Commun. 178:1212-1218(1991).

[2] Otake E., Hashimoto T., Mizuta K.

10 Protein Seq. Data Anal. 5:285-300(1993).

542. Ribosomal protein S14p/S29e (Ribosomal protein S14 signature)

Ribosomal protein S14 is one of the proteins from the small ribosomal subunit. In
 15 Escherichia coli, S14 is known to be required for the assembly of 30S particles and may also
 be responsible for determining the conformation of 16S rRNA at the A site. It belongs to a
 family of ribosomal proteins which, on the basis of sequence similarities [1,2], groups:

- Eubacterial S14.
- Algal and plant chloroplast S14.
- Cyanelle S14.
- Archaeobacterial Methanococcus vannielii S14.
- Plant mitochondrial S14.
- Yeast mitochondrial MRP2.
- Mammalian S29.
- 25 - Yeast YS29A/B.

S14 is a protein of 53 to 115 amino-acid residues. Our signature pattern is based on
 the few conserved positions located in the center of these proteins.

Consensus pattern: [RP]-x(0,1)-C-x(11,12)-[LIVMF]-x-[LIVMF]-[SC]-[RG]-x(3)-[RN]

30

[1] Chan Y.-L., Suzuki K., Olvera J., Wool I.G. Nucleic Acids Res. 21:649-655(1993).

[2] Otake E., Hashimoto T., Mizuta K. Protein Seq. Data Anal. 5:285-300(1993).

543. Ribosomal protein S15 signature

Ribosomal protein S15 is one of the proteins from the small ribosomal subunit.

In *Escherichia coli*, this protein binds to 16S ribosomal RNA and functions at

early steps in ribosome assembly. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1,2], groups: - Eubacterial S15.

- Archaeobacterial *Halobacterium marismortui* HmaS15 (HS11).

- Plant chloroplast S15. - Yeast mitochondrial S28. - Mammalian S13.

- *Brugia pahangi* and *Wuchereria bancrofti* S13 (S15). - Yeast S13 (YS15).

S15 is a protein of 80 to 250 amino-acid residues.

A conserved region located in the C-terminal part of these proteins has been selected as a signature pattern.

-Consensus pattern: [LIVM]-x(2)-H-[LIVMFY]-x(5)-D-x(2)-[SAGN]-x(3)-[LF]-x(9)-[LIVM]-x(2)-[FY]

[1] Dang H., Ellis S.R.

Nucleic Acids Res. 18:6895-6901(1990).

[2] Otake E., Hashimoto T., Mizuta K.

Protein Seq. Data Anal. 5:285-300(1993).

544. Ribosomal protein S16 signature

Ribosomal protein S16 is one of the proteins from the small ribosomal subunit. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups:

- Eubacterial S16.

- Algal and plant chloroplast S16.

- Cyanelle S16.

- *Neurospora crassa* mitochondrial S24 (cyt-21).

S16 is a protein of about 100 amino-acid residues. A conserved region located in the N-terminal extremity of these proteins has been selected as a signature pattern.

Consensus pattern: [LIVMT]-x-[LIVM]-[KR]-L-[STAK]-R-x-G-[AKR]

[1] Otake E., Hashimoto T., Mizuta K. Protein Seq. Data Anal. 5:285-300(1993).

545. Ribosomal protein S17 signature

5 Ribosomal protein S17 is one of the proteins from the small ribosomal subunit.

In *Escherichia coli*, S17 is known to bind specifically to the 5' end of 16S ribosomal RNA and is thought to be involved in the recognition of termination codons. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1,2,3], groups: - Eubacterial S17.

- 10 - Plant chloroplast S17 (nuclear encoded). - Red algal chloroplast S17.
 - Cyanelle S17. - Archaeobacterial S17. - Mammalian and plant cytoplasmic S11.
 - Yeast S18a and S18b (RP41; YS12).

The best conserved regions located in the C-terminal sections of these proteins have been selected as a signature pattern.

5 -Consensus pattern: G-D-x-[LIV]-x-[LIVA]-x-[QEK]-x-[RK]-P-[LIV]-S

[1] Gantt J.S., Thompson M.D. J. Biol. Chem. 265:2763-2767(1990).

[2] Herfurth E., Hirano H., Wittmann-Liebold B.
 Biol. Chem. Hoppe-Seyler 372:955-961(1991).

[3] Otake E., Hashimoto T., Mizuta K.
 20 Protein Seq. Data Anal. 5:285-300(1993).

546. Ribosomal protein S17e signature

A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped

25 on the basis of sequence similarities. One of these families consists of:

- Vertebrates S17 [1]. - *Drosophila* S17 [2]. - *Neurospora crassa* S17 (crp-3).
 - Yeast S17a (RP51A) and S17b (RP51B) [3]. - *Methanococcus jannaschii* MJ0245.

These proteins have from 63 (in archaeobacteria) to 130 to 146 amino acids and are highly conserved. A region in the central part of these proteins has been selected
 30 as a signature.

-Consensus pattern: A-x-I-x-[ST]-K-x-L-R-N-[KR]-I-A-G-[FY]-x-T-H

[1] Chen I.-T., Roufa D.J. Gene 70:107-116(1988).

[2] Maki C., Rhoads D.D., Stewart M.J., van Slyke B., Denell R.E.,

Roufa D.J. Gene 79:289-298(1989).[3] Abovich N., Rosbash M.

Mol. Cell. Biol. 4:1871-1879(1984).

5 547. Ribosomal protein S18 signature

Ribosomal protein S18 is one of the proteins from the small ribosomal subunit. In Escherichia coli, S18 has been involved in aminoacyl-tRNA binding[1]. It appears to be situated at the tRNA A-site of the ribosome. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities[2], groups: - Eubacterial S18. - Algal and plant
10 chloroplast S18. - Cyanelle S18. As a signature pattern, a conserved region in the central section of the protein has been selected. This region contains two basic residues which may be involved in RNA-binding.-

Consensus pattern: [IV]-[DY]-Y-x(2)-[LIVMT]-x(2)-[LIVM]-x(2)-[FYT]-[LIVM]-[ST]-
[DERP]-x-[GY]-K-[LIVM]-x(3)-R-[LIVMAS]-

15 [1] McDougall J., Choli T., Kruft V., Kapp U., Wittmann-Liebold B. FEBS Lett. 245:253-260(1989).[2] Otaka E., Hashimoto T., Mizuta K. Protein Seq. Data Anal. 5:285-300(1993).

20 548. Ribosomal protein S19 signature

Ribosomal protein S19 is one of the proteins from the small ribosomal subunit.

In Escherichia coli, S19 is known to form a complex with S13 that binds strongly to 16S ribosomal RNA. S19 belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1,2], groups: - Eubacterial S19.

- Algal and plant chloroplast S19. - Cyanelle S19. - Archaeobacterial S19.

25 - Plant mitochondrial S19. - Eukaryotic S15 ('rig' protein).

S19 is a protein of 88 to 144 amino-acid residues. Our signature pattern is based on the few conserved positions located in the C-terminal section of these proteins.

-Consensus pattern: [STDNQ]-G-[KMQ]-x(6)-[LIVM]-x(4)-[LIVM]-[GSD]-x(2)-[LF]-
30 [GAS]-[DE]-F-x(2)-[ST]

[1] Kitagawa M., Takasawa S., Kikuchi N., Itoh T., Teraoka H., Yamamoto H., Okamoto H. FEBS Lett. 283:210-214(1991).

[2] Otaka E., Hashimoto T., Mizuta K.

Protein Seq. Data Anal. 5:285-300(1993).

549. Ribosomal protein S19e signature

- 5 A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities [1,2]. One of these families consists of:
- Mammalian S19. - *Drosophila* S19.
 - *Ascaris lumbricoides* S19g (ALEP-1) and S19s. - Yeast YS16 (RP55A and RP55B).
 - *Aspergillus* S16. - *Halobacterium marismortui* HS12.

10 These proteins have 143 to 155 amino acids.

A well conserved stretch of 20 residues in the C-terminal part of these proteins has been selected as a signature pattern.

-Consensus pattern: P-x(6)-[SAN]-x(2)-[LIVMA]-x-R-x-[ALIV]-[LV]-Q-x-L-[EQ]

[1] Etter A., Aboutanos M., Tobler H., Mueller F.

15 Proc. Natl. Acad. Sci. U.S.A. 88:1593-1596(1991).

[2] Suzuki K., Olvera J., Wool I.G. Biochimie 72:299-302(1990).

550. Ribosomal protein S2 signatures

20 Ribosomal protein S2 is one of the proteins from the small ribosomal subunit. S2 belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1,2], groups:

- Eubacterial S2. - Algal and plant chloroplast S2.
- Cyanelle S2. - Archaeobacterial S2.
- Higher eukaryotes P40 (previously thought to be a laminin receptor).
- 25 - Yeast NAB1. - Plant mitochondrial S2. - Yeast mitochondrial MRP4.

S2 is a protein of 235 to 394 amino-acid residues.

Two conserved regions have been selected as signature patterns. One is located in the N-terminal section and the other in the central section.

-Consensus pattern: [LIVMFA]-x(2)-[LIVMFYC](2)-x-[STAC]-[GSTANQEK]-[STALV]-

30 [HY]-[LIVMF]-G

-Consensus pattern: P-x(2)-[LIVMF](2)-[LIVMS]-x-[GDN]-x(3)-[DENL]-x(3)-[LIVM]-x-E-x(4)-[GNQKRH]-[LIVM]-[AP]

[1] Davis S.C., Tzagoloff A., Ellis S.R.

J. Biol. Chem. 267:5508-5514(1992).

[2] Tohgo A., Takasawa S., Munakata H., Yonekura H., Hayashi N., Okamoto H.
FEBS Lett. 340:133-138(1994).

5

551. Ribosomal protein S21 signature

Ribosomal protein S21 is one of the proteins from the small ribosomal subunit. So far S21 has only been found in eubacteria. It is a protein of 55 to 70 amino-acid residues. A conserved region in the N-terminal section of the protein has been selected as a signature pattern.

10

Consensus pattern: [DE]-x-A-[LIY]-[KR]-R-F-K-[KR]-x(3)-[KR]

552. Ribosomal protein S21e signature

A number of eukaryotic ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of: - Mammalian S21 [1].

- Caenorhabditis elegans S21 (F37C12.11). - Rice S21 [2].
- Yeast S21 (Ys25) [3]. - Fission yeast S28 [4].

These proteins have 82 to 87 amino acids.

A perfectly conserved nonapeptide in the N-terminal part of these proteins has been selected as a signature pattern.

-Consensus pattern: L-Y-V-P-R-K-C-S-[SA]

[1] Bhat K.S., Morrison S.G. Nucleic Acids Res. 21:2939-2939(1993).

[2] Nishi R., Hashimoto H., Uchimiya H., Kato A.

Biochim. Biophys. Acta 1216:113-114(1993).[3] Suzuki K., Otake E.

Nucleic Acids Res. 16:6223-6223(1988).[4] Itoh T., Okata E., Matsui K.A.

Biochemistry 24:7418-7423(1985).

25

30 553. Ribosomal protein S24e signature

A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:

- Vertebrate S24 [1]. - Yeast Rp50. - Mucor racemosus S24 [2].

- Halobacterium marismortui HS15 [3]. - Methanococcus jannaschii MJ0394.

These proteins have 101 to 148 amino acids.

A well conserved stretch in the central part of these proteins has been selected as a signature pattern.

5 -Consensus pattern: [FYA]-G-x(2)-[KR]-[STA]-x-G-[FY]-[GA]-x-[LIVM]-Y-[DN]-[SDN]

[1] Brown S.J., Jewell A., Maki C.G., Roufa D.J. Gene 91:293-296(1990).

[2] Sosa L., Fonzi W.A., Sypherd P.S.

Nucleic Acids Res. 17:9319-9331(1989).[3] Kimura J., Arndt E., Kimura M.

10 FEBS Lett. 224:65-70(1987).

554. Ribosomal protein S26e signature

A number of eukaryotic ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of: - Mammalian S26 [1].

- Octopus S26 [2]. - Drosophila S26 (DS31) [3]. - Plant cytoplasmic S26.

- Fungi S26 [4].

These proteins have 114 to 127 amino acids.

A conserved octapeptide in the central part of these proteins has been selected as a signature pattern.

-Consensus pattern: [YH]-C-V-S-C-A-I-H

[1] Kuwano Y., Nakanishi O., Nabeshima Y., Tanaka T., Ogata K.

J. Biochem. 97:983-992(1985).[2] Zinov'eva R.D., Tomarev S.I.

Dokl. Akad. Nauk SSSR 304:464-469(1989).

25 [3] Itoh N., Ohta K., Ohta M., Kawasaki T., Yamashina I.

Nucleic Acids Res. 17:2121-2121(1989).[4] Wu M., Tan H.

Gene 150:401-402(1994).

30 555. Ribosomal protein S28e signature

A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:

- Mammalian S28 [1]. - Plant S28 [2]. - Fungi S33 [3].

- *Methanococcus jannaschii* MJ1202.

These proteins have from 64 to 78 amino acids.

A highly conserved nonapeptide from the C-terminal extremity of these proteins has been selected as a signature pattern.

5 -Consensus pattern: E-[ST]-E-R-E-A-R-x-L

[1] Chan Y.-L., Olvera J., Wool I.G.

Biochem. Biophys. Res. Commun. 179:314-318(1991).

[2] Hwang I., Goodman H.M. Plant Physiol. 102:1357-1358(1993).

[3] Hoekstra R., Ferreira P.M., Bootsman T.C., Mager W.H., Planta R.J.

10 Yeast 8:949-959(1992).

556. Ribosomal protein S3Ae signature

A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:

- Mammalian S3A (was originally known as v-fos transformation effector protein).
- *Caenorhabditis elegans* S3A (F56F3.5).
- Plant cytoplasmic S3A (CYC07) [1].
- Yeast Rp10 (PLC1 and PLC2).
- Fission yeast Rp10 (SpAC13G6.02c).
- *Methanococcus jannaschii* MJ0980.

20 These proteins have from 220 to 250 amino acids.

A conserved stretch in their N-terminal section was selected as a signature pattern.

-Consensus pattern: [LIV]-x-[GH]-R-[IV]-x-E-x-[SC]-L-x-D-L

[1] Liu J.H., Reid D.M.

Plant Physiol. 109:338-338(1995).

25

557. Ribosomal protein S3 signature

Ribosomal protein S3 is one of the proteins from the small ribosomal subunit.

In *Escherichia coli*, S3 is known to be involved in the binding of initiator

30 Met-tRNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups: - Eubacterial S3.

- Algal and plant chloroplast S3. - Cyanelle S3. - Archaeobacterial S3.

- Plant mitochondrial S3. - Vertebrate S3. - Insect S3.

- *Caenorhabditis elegans* S3 (C23G10.3). - Yeast S3 (Rp13).

S3 is a protein of 209 to 559 amino-acid residues.

A conserved region located in the C-terminal section has been selected as a signature pattern.

-Consensus pattern: [GSTA]-[KR]-x(6)-G-x-[LIVMT]-x(2)-[NQSCH]-x(1,3)-[LIVFCA]-
x(3)-[LIV]-[DENQ]-x(7)-[LMT]-x(2)-G-x(2)-G

[1] Otake E., Hashimoto T., Mizuta K.

Protein Seq. Data Anal. 5:285-300(1993).

558. Ribosomal protein S4 signature

Ribosomal protein S4 is one of the proteins from the small ribosomal subunit.

In *Escherichia coli*, S4 is known to bind directly to 16S ribosomal RNA.

Mutations in S4 have been shown to increase translational error frequencies.

It belongs to a family of ribosomal proteins which, on the basis of sequence

similarities [1,2], groups: - Eubacterial S4. - Algal and plant chloroplast S4.

- Cyanelle S4. - Archaeobacterial S4. - Mammalian S9. - Yeast YS11 (SUP45).

- *Marchantia polymorpha* mitochondrial S4. - *Dictyostelium discoideum* rp1024.

- Yeast protein NAM9 [3]. NAM9 has been characterized as a suppressor for ochre mutations in mitochondrial DNA. It could be a ribosomal protein that acts as a suppressor by decreasing translation accuracy.

S4 is a protein of 171 to 205 amino-acid residues (except for NAM9 which is much larger). The signature pattern for this protein is based on a conserved region located in the central section of these proteins.

-Consensus pattern: [LIVM]-[DE]-x-R-[LI]-x(3)-[LIVMC]-[VMFYHQ]-[KRT]-x(3)-
[STAGCVF]-x-[ST]-x(3)-[SAI]-[KR]-x-[LIVMF](2)

[1] Mizuta K., Hashimoto T., Suzuki K.I., Otake E.

Nucleic Acids Res. 19:2603-2608(1991).

[2] Otake E., Hashimoto T., Mizuta K.

Protein Seq. Data Anal. 5:285-300(1993).

[3] Boguta M., Dmochowska A., Borsuk P., Wrobel K., Gargouri A., Lazowska J.,

Slonimski P., Szczesniak B., Kruszewska A.

Mol. Cell. Biol. 12:402-412(1992).

559. Ribosomal protein S4e signature

A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:

- 5 - Mammalian S4 [1]. Two highly similar isoforms of this protein exist : one coded by a gene on chromosome Y, and the other on chromosome X.
 - Plant cytoplasmic S4 [2] - Yeast S7 (YS6). - Archeobacterial S4e.

These proteins have 233 to 264 amino acids.

10 A highly conserved stretch of 15 residues in their N-terminal section has been selected as a signature pattern. Four positions in this region are positively charged residues.

-Consensus pattern: H-x-K-R-[LIVMF]-[SANK]-x-P-x(2)-[WY]-x-[LIVM]-x-[KRP]

[1] Fisher E.M., Beer-Romero P., Brown L.G., Ridley A., McNeil J.A.,
 Lawrence J.B., Willard H.F., Bieber F.R., Page D.C.
 15 Cell 63:1205-1218(1990).

[2] Braun H.P., Emmermann M., Mentzel H., Schmitz U.K.
 Biochim. Biophys. Acta 1218:435-438(1994).

20 560. Ribosomal protein S5 signature

Ribosomal protein S5 is one of the proteins from the small ribosomal subunit.

In *Escherichia coli*, S5 is known to be important in the assembly and function of the 30S ribosomal subunit. Mutations in S5 have been shown to increase translational error frequencies. It belongs to a family of ribosomal proteins

25 which, on the basis of sequence similarities [1,2], groups: - Eubacterial S5.

- Cyanelle S5. - Red algal chloroplast S5. - Archaeobacterial S5.
 - Mammalian S2 (LLrep3). - *Caenorhabditis elegans* S2 (C49H3.11).
 - *Drosophila* S2. - Plant S2. - Yeast S4 (SUP44). - Fungi mitochondrial S5.

30 S5 is a protein of 166 to 254 amino-acid residues. The signature pattern for this protein is based on a conserved region, rich in glycine residues, and located in the N-terminal section of these proteins.

-Consensus pattern: G-[KRQ]-x(3)-[FY]-x-[ACV]-x(2)-[LIVMA]-[LIVM]-[AG]-[DN]-
 x(2)-G-x-[LIVM]-G-x-[SAG]-x(5,6)-[DEQ]-[LIVMA]-x(2)-A-

[LIVMF]

[1] All-Robyn J.A., Brown N., Otaka E., Liebman S.W.

Mol. Cell. Biol. 10:6544-6553(1990).[2] Otaka E., Hashimoto T., Mizuta K.

Protein Seq. Data Anal. 5:285-300(1993).

5

561. Ribosomal protein S6 signature

Ribosomal protein S6 is one of the proteins from the small ribosomal subunit.

In Escherichia coli, S6 is known to bind together with S18 to 16S ribosomal

10 RNA. It belongs to a family of ribosomal proteins which, on the basis of

sequence similarities, groups: - Eubacterial S6. - Red algal chloroplast S6.

- Cyanelle S6.

S6 is a protein of 95 to 208 amino-acid residues. The signature pattern for
this protein is based on a conserved region located in the N-terminal section
of these proteins.

-Consensus pattern: G-x-[KRC]-[DENQRH]-L-[SA]-Y-x-I-[KRNSA]

562. Ribosomal protein S6e signature

A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped
on the basis of sequence similarities. One of these families consists of:

- Mammalian S6 [1]. - Drosophila S6 [2]. - Plant S6 [3]. - Yeast S10 (YS4).

- Halobacterium marismortui HS13 [4]. - Methanococcus jannaschii MJ1260.

S6 is the major substrate of protein kinases in eukaryotic ribosomes [5]; it

25 may have an important role in controlling cell growth and proliferation
through the selective translation of particular classes of mRNA.

These proteins have 135 to 249 amino acids.

A conserved stretch of 12 residues in the N-terminal part of these
proteins has been selected as a signature pattern.

30 -Consensus pattern: [LIVM]-[STAMR]-G-G-x-D-x(2)-G-x-P-M

[1] Franco R., Rosenfeld M.G. J. Biol. Chem. 265:4321-4325(1990).

[2] Watson K.L., Konrad K.D., Woods D.F., Bryant P.J.

Proc. Natl. Acad. Sci. U.S.A. 89:11302-11306(1992).

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
20

- [3] Hansen G., Estruch J.J., Spena A.
Nucleic Acids Res. 20:5230-5230(1992).
- [4] Kimura M., Arndt E., Hatakeyama T., Hatakeyama T., Kimura J.
Can. J. Microbiol. 35:195-199(1989).
- 5 [5] Bandi H.R., Ferrari S., Krieg J., Meyer H.E., Thomas G.
J. Biol. Chem. 268:4530-4533(1993).

563. Ribosomal protein S7 signature

- 10 Ribosomal protein S7 is one of the proteins from the small ribosomal subunit.
In Escherichia coli, S7 is known to bind directly to part of the 3'end of 16S
ribosomal RNA. It belongs to a family of ribosomal proteins which, on the
basis of sequence similarities [1,2,3], groups: - Eubacterial S7.
- Algal and plant chloroplast S7. - Cyanelle S7. - Archaeobacterial S7.
- 5 - Plant mitochondrial S7. - Mammalian S5. - Plant S5.
- Caenorhabditis elegans S5 (T05E11.1).

The best conserved region located in the N-terminal section of these proteins has
been selected as a signature pattern.

-Consensus pattern: [DENSK]-x-[LIVMDET]-x(3)-[LIVMFTA](2)-x(6)-G-K-[KR]-x(5)-
20 [LIVMF]-[LIVMFC]-x(2)-[STAC]

- [1] Klussmann S., Franke P., Bergmann U., Kostka S., Wittmann-Liebold B.
Biol. Chem. Hoppe-Seyler 374:305-312(1993).
- [2] Otake E., Hashimoto T., Mizuta K.
Protein Seq. Data Anal. 5:285-300(1993).
- 25 [3] Ignatovich O., Cooper M., Kulesza H.M., Beggs J.D.
Nucleic Acids Res. 23:4616-4619(1995).

564. Ribosomal protein S7e signature

- 30 A number of eukaryotic ribosomal proteins can be grouped on the basis of sequence
similarities [1]. One of these families consists of:
- Mammalian S7.
- Xenopus S8.

- Insect S7.
- Yeast probable ribosomal protein S7 (N2212).
- Fission yeast probable ribosomal protein S7 (SpAC18G6.13c).

These proteins have about 200 amino acids. A highly conserved stretch of 14 residues which is located in the central section and which is rich in charged residues was selected as a signature pattern.

Consensus pattern: [KR]-L-x-R-E-L-E-K-K-F-[SAP]-x-[KR]-H

[1] Salazar C.E., Mills-Hamm D.M., Kumar V., Collins F.H. Nucleic Acids Res. 21:4147-4147(1993).

565. Ribosomal protein S8 signature

Ribosomal protein S8 is one of the proteins from the small ribosomal subunit.

In *Escherichia coli*, S8 is known to bind directly to 16S ribosomal RNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups:

- Eubacterial S8. - Algal and plant chloroplast S8.
- Cyanelle S8. - Archaeobacterial S8. - *Marchantia polymorpha* mitochondrial S8.
- Mammalian S15A. - Plant S15A. - Yeast S22 (S24).

The best conserved region located in the C-terminal section of these proteins has been selected as a signature pattern.

-Consensus pattern: [GE]-x(2)-[LIV](2)-[STY]-[ST]-x(2)-G-[LIVM](2)-x(4)-[AG]-[KRHAYI]

[1] Otake E., Hashimoto T., Mizuta K.

Protein Seq. Data Anal. 5:285-300(1993).

566. Ribosomal protein S8e signature

A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities [1]. One of these families consists of:

- Mammalian S8. - *Caenorhabditis elegans* S8 (F42C5.8). - *Leishmania major* S8.
- Plant S8. - Yeast S8 (S14) (Rp19). - Archeobacterial S8e.

These proteins have either about 220 amino acids (in eukaryotes) or about 125 amino acids (in archaebacteria). A conserved stretch which is located in the N-terminal section and which is rich in positively charged residues has been selected as a signature pattern.

- 5 -Consensus pattern: [KR]-x(2)-[ST]-G-[GA]-x(5)-[HR]-[KG]-[KR]-x-K-x-E-[LM]-G
 [1] Engemann S., Herfurth E., Briesemeister U., Wittmann-Liebold B.
 J. Protein Chem. 14:189-195(1995).

10 567. Ribosomal protein S9 signature

Ribosomal protein S9 is one of the proteins from the small ribosomal subunit. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1,2], groups: - Eubacterial S9. - Algal chloroplast S9.
 - Cyanelle S9. - Archaeobacterial S9. - Mammalian S16. - Plant S16.
 15 - Yeast mitochondrial ribosomal S9.

A conserved region containing many charged residues and located in the central section of these proteins has been selected as a signature pattern.

-Consensus pattern: G-G-G-x(2)-[GSA]-Q-x(2)-[SA]-x(3)-[GSA]-x-[GSTAV]-[KR]-
 [GSAL]-[LIF]

- 20 [1] Chan Y.-L., Paz V., Olvera J., Wool I.G. FEBS Lett. 263:85-88(1990).
 [2] Otaka E., Hashimoto T., Mizuta K.
 Protein Seq. Data Anal. 5:285-300(1993).

25 568. Ribulose-phosphate 3-epimerase family signatures

Ribulose-phosphate 3-epimerase (EC 5.1.3.1) (also known as pentose-5-phosphate 3-epimerase or PPE) is the enzyme that converts D-ribulose 5-phosphate into D-xylulose 5-phosphate in Calvin's reductive pentose phosphate cycle. In *Alcaligenes eutrophus* two copies of the gene coding for PPE are known [1],
 30 one is chromosomally encoded (cbbEC), the other one is on a plasmid (cbbEP). PPE has been found in a wide range of bacteria, archaebacteria, fungi and plants. The sequence of PPE is highly related to:

- *Escherichia coli* D-allulose-6-phosphate 3-epimerase (gene *alsE*).

- Escherichia coli protein sgcE.
- Mycoplasma genitalium hypothetical protein MG112.

All these proteins have from 209 to 241 amino acid residues.

Two conserved regions which are located respectively in the N-terminal and in the central part of these proteins have been selected as signature patterns.

-Consensus pattern: [LIVMF]-H-[LIVMFY]-D-[LIVM]-x-D-x(1,2)-[FY]-[LIVM]-x-N-x-[STAV]

-Consensus pattern: [LIVMA]-x-[LIVM]-M-[ST]-[VS]-x-P-x(3)-G-Q-x-F-x(6)-[NK]-[LIVMC]

[1] Kusian B., Yoo J.G., Bednarski R., Bowien B.
J. Bacteriol. 174:7337-7344(1992).

569. (Ricin B lectin) Similarity to lectin domain of ricin beta-chain, 3 copies.

This family consists of a triplicated domain involved in cell agglutination in ricin.

570. (Rotamase) PpiC-type peptidyl-prolyl cis-trans isomerase signature
Peptidyl-prolyl cis-trans isomerase (EC 5.2.1.8) (PPIase or rotamase) is an enzyme that accelerates protein folding by catalyzing the cis-trans isomerization of proline imidic peptide bonds in oligopeptides [1]. Most characterized PPIases belong to two families, the cyclophilin-type (see <PDOC00154>) and the FKBP-type (see <PDOC00426>). Recently a third family has been discovered [2,3]. So far, the only biochemically characterized member of this family is the Escherichia coli protein parvulin (gene ppiC), a small (92 residues) cytoplasmic enzyme that prefers amino acid residues with hydrophobic side chains like leucine and phenylalanine in the P1 position of the peptides substrates. PpiC is evolutionary related to a number of proteins that are also probably PPIases:

- Escherichia coli and Haemophilus influenzae ppiD. PpiD is a PPIase which contains a periplasmic ppiC-like domain anchored to the inner membrane and

which seems to be involved in the folding of outer membrane proteins.

- *Escherichia coli* surA. SurA is a periplasmic protein that contains two ppiC-like domains.
- Nitrogen-assimilating bacteria protein nifM which is involved in the activation and stabilization of the iron-component (nifH) of nitrogenase.
- *Bacillus subtilis* protein prsA, a membrane-bound lipoprotein involved in protein export.
- *Lactococcus* and *lactobacillus* protease maturation protein prtM, a membrane-bound lipoprotein involved in the maturation of a secreted serine proteinase.
- Yeast protein ESS1/PTF1 (processing/termination factor 1).
- *Drosophila* protein dodo (gene dod).
- Mammalian protein PIN1,
- *Campylobacter jejuni* cell binding factor 2 (CBF2), a secreted antigen.
- *Bacillus subtilis* hypothetical protein yacD.
- *Helicobacter pylori* hypothetical protein HP0175.
- A hypothetical slime mold protein.

A conserved region that contains a serine which could play a role in the catalytic mechanism of these enzymes has been selected as a signature pattern.

-Consensus pattern: F-[GSADEI]-x-[LVAQ]-A-x(3)-[ST]-x(3,4)-[STQ]-x(3,5)-[GER]-G-x-[LIVM]-[GS]

[1] Fischer G., Schmid F.X.

Biochemistry 29:2205-2212(1990).

[2] Rudd K.E., Sofia H.J., Koonin E.V., Plunkett G. III, Lazar S.,

Rouviere P.E. Trends Biochem. Sci. 20:14-15(1995).

[3] Rahfeld J.-U., Ruecknagel K.P., Schelbert B., Ludwig B., Hacker J.,

Mann K., Fischer G. FEBS Lett. 352:180-184(1994).

571. (RrnaAD) Ribosomal RNA adenine dimethylases signature

A number of enzymes responsible for the dimethylation of adenosines in ribosomal RNAs (EC 2.1.1.48) have been found [1,2] to be evolutionary related.

These enzymes are:

- Bacterial 16S rRNA dimethylase (gene ksgA), which acts in the biogenesis of ribosomes by catalyzing the dimethylation of two adjacent adenosines in

the loop of a conserved hairpin near the 3'-end of 16S rRNA. Inactivation of ksgA leads to resistance to the aminoglycoside antibiotic kasugamycin.

- Yeast 18S rRNA dimethylase (gene DIM1), which is functionally similar to ksgA and that dimethylates twin adenosines in the 3'-end of 18S rRNA.

- 5 - Bacterial 'erm' methylases. These enzymes confer resistance to macrolide-lincosamide-streptogramin B (MLS) antibiotics - such as erythromycin - by dimethylating the adenine residue at position 2058 of 23S rRNA thus resulting in a reduced affinity between ribosomes and the MLS antibiotics.
- *Caenorhabditis elegans* hypothetical protein EO2H1.1.

10 The best conserved regions in these enzymes is located in the N-terminal section and corresponds to a region that is probably involved in S-adenosyl methionine (SAM) binding.

-Consensus pattern: [LIVM]-[LIVMFY]-[DE]-x-G-[STAPV]-G-x-[GA]-x-[LIVMF]-[ST]-x(2)-[LIVM]-x(6)-[LIVMY]-x-[STAGV]-[LIVMFYHC]-E-x-D

15 [1] van Gemen B., van Knippenberg P.H.

(In) Nucleic acid methylation, Clawson G.A., Willis D.B., Weissbach A., Jones P.A., Eds., pp.19-36, Alan R. Liss Inc, New-York, (1990).

[2] Lafontaine D., Delcour J., Glasser A.L., Desgres J., Vandenhoute J. J. Mol. Biol. 241:492-497(1994).

572. (RuBisC0 small) Ribulose bisphosphate carboxylase, small chain. 206 members

25 573. ATP/GTP-binding site motif A (P-loop) (ras)

From sequence comparisons and crystallographic data analysis it has been shown [1,2,3,4,5,6] that an appreciable proportion of proteins that bind ATP or GTP share a number of more or less conserved sequence motifs. The best conserved of these motifs is a glycine-rich region, which typically forms a flexible loop between a beta-strand and an alpha-helix.

30 This loop interacts with one of the phosphate groups of the nucleotide. This sequence motif is generally referred to as the 'A' consensus sequence [1] or the 'P-loop' [5]. There are numerous ATP- or GTP-binding proteins in which the P-loop is found. A number of protein families for which the relevance of the presence of such a motif has been noted are listed below: - ATP

synthase alpha and beta subunits. - Myosin heavy chains. - Kinesin heavy chains and kinesin-like proteins. - Dynamins and dynamin-like proteins - Guanylate kinase - Thymidine kinase (-Thymidylate kinase. - Shikimate kinase. - Nitrogenase iron protein family (nifH/frxC) - ATP-binding proteins involved in 'active transport' (ABC transporters) [7] - DNA and RNA

5 helicases [8,9,10]. - GTP-binding elongation factors (EF-Tu, EF-1alpha, EF-G, EF-2, etc.). - Ras family of GTP-binding proteins (Ras, Rho, Rab, Ral, Ypt1, SEC4, etc.). - Nuclear protein ran. - ADP-ribosylation factors family - Bacterial dnaA protein - Bacterial recA protein - Bacterial recF protein - Guanine nucleotide-binding proteins alpha subunits (Gi, Gs, Gt, G0, etc.). - DNA mismatch repair proteins mutS family - Bacterial type II secretion system

10 protein E. Not all ATP- or GTP-binding proteins are picked-up by this motif. A number of proteins escape detection because the structure of their ATP-binding site is completely different from that of the P-loop. Examples of such proteins are the E1-E2 ATPases or the glycolytic kinases. In other ATP- or GTP-binding proteins the flexible loop exists in a slightly different form; this is the case for tubulins or protein kinases. A special mention must be reserved for adenylate kinase, in which there is a single deviation from the P-loop pattern: in the last position Gly is found instead of Ser or Thr.

Consensus pattern: [AG]-x(4)-G-K-[ST]

In addition to the proteins listed above, the 'A' motif is also found in a number of other proteins. Most of these proteins probably bind a nucleotide, but others are definitively not ATP- or GTP-binding (as for example chymotrypsin, or human ferritin light chain).

[1] Walker J.E., Saraste M., Runswick M.J., Gay N.J. EMBO J. 1:945-951(1982).[2] Moller W., Amons R. FEBS Lett. 186:1-7(1985).[3] Fry D.C., Kuby S.A., Mildvan A.S. Proc. Natl. Acad. Sci. U.S.A. 83:907-911(1986).[4] Dever T.E., Glynias M.J., Merrick W.C. Proc. Natl. Acad. Sci. U.S.A. 84:1814-1818(1987).[5] Saraste M., Sibbald P.R., Wittinghofer A. Trends Biochem. Sci. 15:430-434(1990).[6] Koonin E.V. J. Mol. Biol. 229:1165-1174(1993).[7] Higgins C.F., Hyde S.C., Mimmack M.M., Gileadi U., Gill D.R., Gallagher M.P. J. Bioenerg. Biomembr. 22:571-592(1990).[8] Hodgman T.C. Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).[9] Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K., Schnier J., Slonimski P.P. Nature 337:121-122(1989).[10] Gorbalenya A.E., Koonin E.V., Donchenko A.P., Blinov V.M. Nucleic Acids Res. 17:4713-4730(1989).

GTP-binding nuclear protein ran signature (ras)

Ran (or TC4) is a small abundant nuclear protein that binds and hydrolyzes GTP and which has been implicated in a large number of processes including nucleocytoplasmic transport, RNA synthesis, processing and export and cell cycle checkpoint control [1,2]. Ran is generally included in the RAS 'superfamily' of small GTP-binding proteins [3], but it is only slightly related to the other RAS proteins. It also differs from RAS proteins in that it lacks cysteine residues at its C- terminal and is therefore not subject to prenylation. Instead ran has an acidic C-terminus. It is, however similar to RAS family members in requiring a specific guanine nucleotide exchange factor (GEF) and a specific GTPase activating protein (GAP) as stimulators of overall GTPase activity. The region of the GTP-binding B motif which, in ran, is perfectly conserved has been selected as a signature pattern.

Consensus pattern: D-T-A-G-Q-E-K-[LF]-G-G-L-R-[DE]-G-Y-Y- Proteins belonging to this family also contain a copy of the ATP/GTP- binding motif 'A' (P-loop).

[1] Scheffzek K., Klebe C., Fritz-Wolf K., Kabsch W., Wittinghofer A. Nature 374:378-

381(1995).[2] Rush M.G., Drivas G., d'Eustachio P. BioEssays 18:103-112(1996).[3]

Valencia A., Chardin P., Wittinghofer A., Sander C. Biochemistry 30:4637-4648(1991).

574. recA signature

The bacterial recA protein [1,2,3,E1] is essential for homologous recombination and recombinational repair of DNA damage. RecA has many activities: it filaments, it binds to single- and double-stranded DNA, it binds and hydrolyzes ATP, it is also a recombinase and, finally, it interacts with lexA causing its activation and leading to its autocatalytic cleavage. RecA is a protein of about 350 amino-acid residues. Its sequence is very well conserved [3,4,5,E1] among eubacterial species. It is also found in the chloroplast of plants [6]. The best conserved region, a nonapeptide located in the middle of the sequence which is part of the monomer-monomer interface in a recA filament has been selected as a signature pattern,.

Consensus pattern: A-L-[KR]-[IF]-[FY]-[STA]-[STAD]-[LIVMQ]-R-

[1] Smith K.C., Wang T.-C. V. BioEssays 10:12-16(1989).[2] Lloyd A.T., Sharp P.M. J.

Mol. Evol. 37:399-407(1993).[3] Roca A.I., Cox M.M. Prog. Nucleic Acids Res. Mol. Biol.

56:129-223(1997).[4] Karlin S., Weinstock G.M., Brendel V. J. Bacteriol. 177:6881-

6893(1995).[5] Eisen J.A. J. Mol. Evol. 41:1105-1123(1995).[6] Cerutti H.D., Osman M.,

Grandoni P., Jagendorf A.T. Proc. Natl. Acad. Sci. U.S.A. 89:8068-8072(1992).[E1]

<http://www.tigr.org/~jeisen/RecA/RecA.html>

575. Response regulator receiver domain

This domain receives the signal from the sensor partner inComment: bacterial two-
 5 component systems. It is usually found N-terminalComment: to a DNA binding effector domain.

[1] Pao GM, Saier MH; J Mol Evol 1995;40:136-154.

10 576. Ribonucleotide reductase large subunit signature

*Ribonucleotide reductase (EC 1.17.4.1) [1,2] catalyzes the reductive synthesis of
 deoxyribonucleotides from their corresponding ribonucleotides. It provides the precursors
 necessary for DNA synthesis. Ribonucleotide reductase is an oligomeric enzyme composed
 of a large subunit (700 to 1000 residues) and a small subunit (300 to 400 residues). There are
 15 regions of similarities in the sequence of the large chain from prokaryotes, eukaryotes and viruses. One of these regions has been selected as a signature pattern.

Consensus pattern: W-x(2)-[LF]-x(6,7)-G-[LIVM]-[FYRA]-[NH]-x(3)-[STAQLIVM]-
 [ASC]-x(2)-[PA]-

[1] Nillson O., Lundqvist T., Hahne S., Sjoberg B.-M. Biochem. Soc. Trans. 16:91-
 20 94(1988).[2] Reichard P. Science 260:1773-1777(1993).

577. Ribonuclease T2 family histidine active sites

The fungal ribonucleases T2 from *Aspergillus oryzae*, M from *Aspergillus saitoi* and Rh from
 25 *Rhizopus niveus* are structurally and functionally related 30 Kd glycoproteins [1] that cleave the 3'-5' internucleotide linkage of RNA via a nucleotide 2',3'-cyclic phosphate intermediates (EC 3.1.27.1). A number of other RNAses have been found to be evolutionary related to these fungal enzymes: - Self-incompatibility [2] in flowering plants is often controlled by a single gene (S-gene) that has several alleles. This gene prevents fertilization by self-pollen or by
 30 pollen bearing either of the two S- alleles expressed in the style. The self-incompatibility glycoprotein from several higher plants of the solanaceae family has been shown [2,3] to be a ribonuclease. - Phosphate-starvation induced RNAses LE and LX from tomato [4]. These two enzymes are probably involved in a phosphate-starvation rescue system. - *Escherichia coli*

periplasmic RNase I (EC 3.1.27.6) (gene rna) [5]. - *Aeromonas hydrophila* periplasmic RNase. - *Haemophilus influenzae* hypothetical protein HI0526. Two histidines residues have been shown [6,7] to be involved in the catalytic mechanism of RNase T2 and Rh. These residues and the region around them are highly conserved in all the sequence described above.

Two signature patterns have been developed, one for each of the two active-site histidines. The second pattern also contains a cysteine which is known to be involved in a disulfide bond.

Consensus pattern: [FYWL]-x-[LIVM]-H-G-L-W-P [H is an active site residue]

Consensus pattern: [LIVMF]-x(2)-[HDGTY]-[EQ]-[FYW]-x-[KR]-H-G-x-C [H is an active site residue] [C is involved in a disulfide bond]

[1] Watanabe H., Naitoh A., Suyama Y., Inokuchi N., Shimada H., Koyama T., Ohgi K., Irie M. J. Biochem. 108:303-310(1990).[2] Haring V., Gray J.E., McClure B.A., Anderson M.A., Clarke A.E. Science 250:937-941(1990).[3] McClure B.A., Haring V., Ebert P.R., Anderson M.A., Simpson R.J., Sakiyama F., Clarke A.E. Nature 342:95957(1989).[4] Loeffler A., Glund K., Irie M. Eur. J. Biochem. 214:627-633(1993).[5] Meador J. III, Kennell D. Gene 95:1-7(1990).[6] Kawata Y., Sakiyama F., Hayashi F., Kyogoku Y. Eur. J. Biochem. 187:255-262(1990).[7] Kurihara H., Mitsui Y., Ohgi K., Irie M., Mizuno H., Nakamura K.T. FEBS Lett. 306:189-192(1992).

578. Ribonucleotide reductase large subunit signature. Ribonucleotide reductase (EC 1.17.4.1) [1,2] catalyzes the reductive synthesis of deoxyribonucleotides from their corresponding ribonucleotides. It provides the precursors necessary for DNA synthesis. Ribonucleotide reductase is an oligomeric enzyme composed of a large subunit (700 to 1000 residues) and a small subunit (300 to 400 residues). There are regions of similarities in the sequence of the large chain from prokaryotes, eukaryotes and viruses. One of these regions has been developed as a signature pattern.

Consensus pattern: W-x(2)-[LF]-x(6,7)-G-[LIVM]-[FYRA]-[NH]-x(3)-[STAQLIVM]-[ASC]-x(2)-[PA]-

[1] Nilsson O., Lundqvist T., Hahne S., Sjöberg B.-M. Biochem. Soc. Trans. 16:91-94(1988).[2] Reichard P. Science 260:1773-1777(1993).

579. RNase H

RNase H digests the RNA strand of an RNA/DNA hybrid. Important enzyme in retroviral replication cycle, and often found as a domain associated with reverse transcriptases.

Structure is a mixed alpha+beta fold with three a/b/a layers.

580. Eukaryotic putative RNA-binding region RNP-1 signature (rrm)

Many eukaryotic proteins that are known or supposed to bind single-stranded RNA contain one or more copies of a putative RNA-binding domain of about 90 amino acids [1,2]. This region has been found in the following proteins: ** Heterogeneous nuclear ribonucleoproteins ** - hnRNP A1 (helix destabilizing protein) (twice). - hnRNP A2/B1 (twice). - hnRNP C (C1/C2) (once). - hnRNP E (UP2) (at least once). - hnRNP G (once). ** Small nuclear ribonucleoproteins ** - U1 snRNP 70 Kd (once). - U1 snRNP A (once). - U2 snRNP B'' (once). ** Pre-RNA and mRNA associated proteins ** - Protein synthesis initiation factor 4B (eIF-4B) [3], a protein essential for the binding of mRNA to ribosomes (once). - Nucleolin (4 times). - Yeast single-stranded nucleic acid-binding protein (gene SSB1) (once). - Yeast protein NSR1 (twice). NSR1 is involved in pre-rRNA processing; it specifically binds nuclear localization sequences. - Poly(A) binding protein (PABP) (4 times). ** Others ** - Drosophila sex determination protein Sex-lethal (Sxl) (twice). - Drosophila sex determination protein Transformer-2 (Tra-2) (once). - Drosophila 'elav' protein (3 times), which is probably involved in the RNA metabolism of neurons. - Human paraneoplastic encephalomyelitis antigen HuD (3 times) [4], which is highly similar to elav and which may play a role in neuron-specific RNA processing. - Drosophila 'bicoid' protein (once) [5], a segment-polarity homeobox protein that may also bind to specific mRNAs. - La antigen (once), a protein which may play a role in the transcription of RNA polymerase III. - The 60 Kd Ro protein (once), a putative RNP complex protein. - A maize protein induced by abscisic acid in response to water stress, which seems to be a RNA-binding protein. - Three tobacco proteins, located in the chloroplast [6], which may be involved in splicing and/or processing of chloroplast RNAs (twice). - X16 [7], a mammalian protein which may be involved in RNA processing in relation with cellular proliferation and/or maturation. - Insulin-induced growth response protein Cl-4 from rat (twice). - Nucleolysins TIA-1 and

TIAR (3 times) [8] which possesses nucleolytic activity against cytotoxic lymphocyte target cells. may be involved in apoptosis. - Yeast RNA15 protein, which plays a role in mRNA stability and/or poly-(A) tail length [9]. Inside the putative RNA-binding domain there are two regions which are highly conserved. The first one is a hydrophobic segment of six residues (which is called the RNP-2 motif), the second one is an octapeptide motif (which is called RNP-1 or RNP-CS). The position of both motifs in the domain is shown in the following schematic representation:

xxxxxxxx#####xx#####xx

RNP-2 RNP-1

The RNP-1 motif has been used as a signature pattern for this type of domain.

Consensus pattern: [RK]-G-{EDRKHPCG}-[AGSCI]-[FY]-[LIVA]-x-[FYLM] In most cases the residue in position 3 of the pattern is either Tyr or Phe.

[1] Bandziulis R.J., Swanson M.S., Dreyfuss G. *Genes Dev.* 3:431-437(1989).[2] Dreyfuss G., Swanson M.S., Pinol-Roma S. *Trends Biochem. Sci.* 13:86-91(1988).[3] Milburn S.C., Hershey J.W.B., Davies M.V., Kelleher K., Kaufman R.J. *EMBO J.* 9:2783-2790(1990).[4] Szabo A., Dalmau J., Manley G., Rosenfeld M., Wong E., Henson J., Posner J.B., Furneaux H.M. *Cell* 67:325-333(1991).[5] Rebagliati M. *Cell* 58:231-232(1989).[6] Li Y., Sugiura M. *EMBO J.* 9:3059-3066(1990).[7] Ayane M., Preuss U., Koehler G., Nielsen P.J. *Nucleic Acids Res.* 19:1273-1278(1991).[8] Kawakami A., Tian Q., Duan X., Streuli M., Schlossman S.F., Anderson P. *Proc. Natl. Acad. Sci. U.S.A.* 89:8681-8685(1992).[9] Minvielle-Sebastia L., Winsor B., Bonneaud N., Lacroute F. *Mol. Cell. Biol.* 11:3075-3087(1991).

581. Rubredoxin signature

Rubredoxins [1] are small electron-transfer prokaryotic proteins. They contain an iron atom which is ligated by four cysteine residues. Rubredoxins are, in some cases, functionally interchangeable with ferredoxins.

A conserved region that includes two of the cysteine residues that bind the iron atom has been selected as a pattern for these proteins.

Consensus pattern: [LIVM]-x(3)-W-x-C-P-x-C-[AGD] [The two C's bind the iron atom]

In *Pseudomonas oleovorans* rubredoxin 2 (gene *alkG*) [2], this pattern is found twice because *alkG* has two rubredoxin domains.

Rubrerhythrin [3], a protein with inorganic pyrophosphatase activity from *Desulfovibrio vulgaris* possesses a C-terminal rubredoxin-like domain, but this domain is too divergent to be detected by the above pattern.

[1] Berg J.M., Holm R.H.(In) Iron-sulfur proteins, Spiro T.G., Ed., pp1-66, Wiley, New-York, (1982). [2] Kok M., Oldenhuis R., der Linden M.P.G., Meulenberg C.H.C., Kingma J., Witholt B., J. Biol. Chem. 264:5442-5451(1989). [3] van Beeumen J.J., van Driessche G., Liu M.-Y., Le Gall J., J. Biol. Chem. 266:20645-20653(1991).

582. (rvp) Eukaryotic and viral aspartyl proteases active site

Aspartyl proteases, also known as acid proteases, (EC 3.4.23.-) are a widely distributed family of proteolytic enzymes [1,2,3] known to exist invertebrates, fungi, plants, retroviruses and some plant viruses. Aspartate proteases of eukaryotes are monomeric enzymes which consist of two domains. Each domain contains an active site centered on a catalytic aspartyl residue. The two domains most probably evolved from the duplication of an ancestral gene encoding a primordial domain. Currently known eukaryotic aspartyl proteases are: -

Vertebrate gastric pepsins A and C (also known as gastricsin). - Vertebrate chymosin (rennin), involved in digestion and used for making cheese. - Vertebrate lysosomal cathepsins D (EC 3.4.23.5) and E (EC 3.4.23.34). - Mammalian renin (EC 3.4.23.15) whose function is to generate angiotensin I from angiotensinogen in the plasma. - Fungal proteases such as aspergillopepsin A (EC 3.4.23.18), candidapepsin (EC 3.4.23.24), mucoropepsin (EC 3.4.23.23) (mucor rennin), endothiapepsin (EC 3.4.23.22), polyporopepsin (EC 3.4.23.29), and rhizopuspepsin (EC 3.4.23.21). - Yeast saccharopepsin (EC 3.4.23.25) (proteinase A) (gene PEP4). PEP4 is implicated in posttranslational regulation of vacuolar hydrolases. - Yeast barrier pepsin (EC 3.4.23.35) (gene BAR1); a protease that cleaves alpha-factor and thus acts as an antagonist of the mating pheromone. - Fission yeast *sxa1* which is involved in degrading or processing the mating pheromones. Most retroviruses and some plant viruses, such as badnaviruses, encode for anaspartyl protease which is an homodimer of a chain of about 95 to 125 amino acids. In most retroviruses, the protease is encoded as a segment of a polyprotein which is cleaved during the maturation process of the virus. It is generally part of the pol polyprotein and, more rarely, of the gagpolyprotein. Conservation of the sequence

around the two aspartates of eukaryotic aspartyl proteases and around the single active site of the viral proteases allows us to develop a single signature pattern for both groups of protease.

Consensus pattern: [LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-
[STAPDENQ]-x-[LIVMFSTNC]-x-[LIVMFGTA] [D is the active site residue] –

- 5 [1] Foltmann B. Essays Biochem. 17:52-84(1981).[2] Davies D.R. Annu. Rev. Biophys. Chem. 19:189-215(1990).[3] Rao J.K.M., Erickson J.W., Wlodawer A. Biochemistry 30:4663-4671(1991).[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:105-120(1995).

10 583. (rvt) Reverse transcriptase (RNA-dependent DNA polymerase)

A reverse transcriptase gene is usually indicative of a mobile element such as a retrotransposon or retrovirus. Reverse transcriptases occur in a variety of mobile elements, including retrotransposons, retroviruses, group II introns, bacterial msDNAs, hepadnaviruses, and caulimoviruses. Number of members: 1233

[1] Medline: 91006031. Origin and evolution of retroelements based upon their reverse transcriptase sequences. Xiong Y, Eickbush TH; EMBO J 1990;9:3353-3362.

20 584. (S-AdoMet synt) S-adenosylmethionine synthetase signatures

S-adenosylmethionine synthetase (EC 2.5.1.6) is the enzyme that catalyzes the formation of S-adenosylmethionine (AdoMet) from methionine and ATP [1]. AdoMet is an important methyl donor for transmethylation and is also the propylamino donor in polyamine biosynthesis. In bacteria there is a single isoform of AdoMet synthetase (gene metK), there are two in budding yeast (genes SAM1 and SAM2) and in mammals while in plants there is generally a multigene family. The sequence of AdoMet synthetase is highly conserved throughout isozymes and species. Two signature patterns have been selected for this type of enzyme; the first is a hexapeptide which seems to be involved in ATP-binding; the second is an almost perfectly conserved glycine-rich nonapeptide.

25 Consensus pattern: G-A-G-D-Q-G-x(3)-G-[FYH]-Sequences known to belong to this class detected by the pattern:

Consensus pattern: G-[GA]-G-[ASC]-F-S-x-K-[DE]

[1] Horikawa S., Sasuga J., Shimizu K., Ozasa H., Tsukada K. J. Biol. Chem. 265:13683-13686(1990).

5 585. S1 RNA binding domain

The S1 domain occurs in a wide range of RNAComment: associated proteins. It is structurally similarComment: to cold shock protein which binds nucleic acids.Comment: The S1 domain has an OB-fold structure.

[1] Bycroft M, Hubbard TJ, Proctor M, Freund SM, Murzin AG; Cell 1997;88:235-242.

10

586. SAICAR synthetase signatures

Phosphoribosylaminoimidazole-succinocarboxamide synthase (EC 6.3.2.6)

(SAICARsynthetase) catalyzes the seventh step in the de novo purine biosynthetic pathway; the ATP-dependent conversion of 5'-phosphoribosyl-5-aminoimidazole-4-carboxylic acid and aspartic acid to SAICAR [1]. In bacteria (gene purC),fungi (gene ADE1) and plants, SAICAR synthetase is a monofunctional protein;in higher vertebrates it is the N-terminal domain of a bifunctional enzyme that also catalyze phosphoribosylaminoimidazole carboxylase (AIRC) activity. Two conserved regions in the central section of this enzyme have been selected as signature patterns for SAICAR synthetase.

Consensus pattern: [LIVMF](2)-P-[LIVM]-E-x-[LIVM]-[LIVMCA]-R-x(3)-[TA]-G-S-

Consensus pattern: [LIVM]-[LIVMA]-D-x-K-[LIVMFY]-E-F-G

[1] Zalkin H., Dixon J.E. Prog. Nucleic Acid Res. Mol. Biol. 42:259-287(1992).

25

587. (SCP) Extracellular proteins SCP/Tpx-1/Ag5/PR-1/Sc7 signatures

A variety of extracellular proteins from eukaryotes have been found to be evolutionary related: - Rodent sperm-coating glycoprotein (SCP), also known as acidic epididymal glycoprotein (AEG) . This protein is thought to be involved in sperm maturation [1]. It is a protein of about 220 residues and probably contains eight disulfide bonds. - Mammalian testis-specific protein Tpx-1 [2]. Tpx-1 is highly related to SCP's. - Mammalian glioma pathogenesis-related protein (GliPR). - Lizard heliothermine, a toxin that blocks ryanodine receptors. - Venom allergen 5 (Ag5) from vespid wasps and venom allergen 3 (Ag3) from

30

fire ants. These proteins are potent allergens and are the main cause of allergic reactions to stings from insects of the hymenoptera family [3]. Ag5/3 are proteins of about 200 residues and contain four disulfide bonds. - Plant pathogenesis proteins of the PR-1 family [4]. These proteins are synthesized during pathogen infection or other stress-related responses. They are

5 proteins of about 130 to 140 residues and probably contain three disulfide bonds. - Proteins Sc7 and Sc14 from the basidiomycete fungus *Schizophyllum commune*. These extracellular proteins are loosely associated with fruit body hyphal walls [5]. Sc7/14 are proteins of about 180 residues and probably contain two disulfide bonds. - *Ancylostoma* secreted protein from dog hookworm. - Yeast hypothetical proteins YJL078c, YJL079c and YKR013w. The exact

10 function of these proteins is not yet known. Two conserved regions located in their C-terminal half have been selected as signature patterns. The second signature contains a cysteine which is known to be involved in a disulfide bond in Ag5.

Consensus pattern: [GDER]-H-[FYWH]-T-Q-[LIVM](2)-W-x(2)-[STN]

Consensus pattern: [LIVMFYH]-[LIVMFY]-x-C-[NQRHS]-Y-x-[PARH]-x-[GL]-N-[LIVMFYWDN] [C is involved in a disulfide bond]

[1] Mizuki N., Kasahara M. *Mol. Cell. Endocrinol.* 89:25-32(1992).[2] Kasahara M., Gutknecht J., Brew K., Spurr N., Goodfellow P.N. *Genomics* 5:527-534(1989).[3] Lu G., Villalba M., Coscia M.R., Hoffman D.R., King T.P. *J. Immunol.* 150:2823-2830(1993).[4] Dixon D.C., Cutt J.R., Klessig D.F. *EMBO J.* 10:1317-1324(1991).[5] Schuren F.H.J., Asgeirsdottir S.A., Kothe E.M., Scheer J.M.J., Wessels J.G.H. *J. Gen. Microbiol.* 139:2083-2090(1993).

588. SET domain

25 SET domains appear to be protein-protein interaction domains. It has been demonstrated that SET domains mediate interactions with a family of proteins that display similarity with dual-specificity phosphatases (dsPTases) [2].

[1] Tripoulas N, LaJeunesse D, Gildea J, Shearn A; *Genetics* 1996;143:913-928. [2] Cui X, De Vivo I, Slany R, Miyamoto A, Firestein R, Cleary, ML; *Nat Genet* 1998;18:331-337.

30

589. Src homology 3 (SH3) domain profile

The Src homology 3 (SH3) domain is a small protein domain of about 60 amino-acid residues first identified as a conserved sequence in the non-catalytic part of several cytoplasmic protein tyrosine kinases (e.g. Src, Abl, Lck) [1]. Since then, it has been found in a great variety of other intracellular or membrane-associated proteins [2,3,4,5]. The SH3 domain has a characteristic fold which consists of five or six beta-strands arranged as two tightly packed anti-parallel beta sheets. The linker regions may contain short helices [6]. The function of the SH3 domain is not well understood. The current opinion is that they mediate assembly of specific protein complexes via binding to proline-rich peptides [7]. In general SH3 domains are found as single copies in a given protein, but there is a significant number of protein with two SH3 domains and a few with 3 or 4 copies. So far, SH3 domains have been identified in the following proteins: - Many vertebrate, invertebrate and retroviral cytoplasmic (non-receptor) protein tyrosine kinases. In particular in the Src, Abl, Bkt, Csk and ZAP70 families of kinases. - Mammalian phosphatidylinositol-specific phospholipase C-gamma-1 and -2. - Mammalian phosphatidyl inositol 3-kinase regulatory p85 subunit. - Mammalian Ras GTPase-activating protein (GAP). - Adaptor proteins mediating binding of guanine nucleotide exchange factors to growth factor receptors: vertebrate GRB2, *Caenorhabditis elegans* sem-5 and *Drosophila* DRK. All of which have two SH3 domains. - Mammalian Vav oncoprotein, a guanine nucleotide exchange factor of the CDC24 family. - Some guanine-nucleotide releasing factors of the CDC25 family: yeast CDC25, yeast SCD25, fission yeast ste6. - MAGUK proteins. These proteins consist of at least three types of domains: one or more copies of the DHR domain, a SH3 domain and a C-terminal guanylate kinase domain. Members of this family are: *Drosophila* lethal(1) discs large-1 tumor suppressor protein (gene Dlg1), mammalian tight junction protein ZO-1, vertebrate erythrocyte membrane protein p55, *Caenorhabditis elegans* protein lin-2, rat protein CASK and mammalian synaptic proteins SAP90/PSD-95, CHAPSYN-110/PSD-93, SAP97/DLG1 and SAP102. - Miscellaneous proteins interacting with vertebrate receptor protein tyrosine kinases: mammalian cytoplasmic protein Nck (3 copies), oncoprotein Crk (2 copies). - Chicken Src substrate p80/85 protein (cortactin) and the similar human hemopoietic lineage cell specific protein Hs1. - Mammalian dihydropyridine-sensitive L-type calcium channel beta (regulatory) subunit including the related human myasthenic syndrome antigen B (MSYB). - Mammalian neutrophil cytosolic activators of NADPH oxidase: p47 (NCF-1), p67 (NCF-2), and a potential homolog from *Caenorhabditis elegans* (B0303.7). NCF-1 and -2 have two copies of the SH3 domain, while B0303.7 has four. - Some myosin heavy chains from amoebae, slime

molds and yeast (gene MYO3). - Vertebrate and Drosophila spectrin and fodrin alpha-chain. - Human amphiphysin. - Yeast actin-binding protein ABP1. - Yeast actin-binding protein SLA1 (3 copies). - Yeast protein BEM1 and the fission yeast homolog scd2 (or ral3) (2 copies). - Yeast BEM1-binding proteins BOI2 (BEB1) and BOB1 (BOI1). - Yeast fusion protein FUS1. - Yeast protein RSV167. - Yeast protein SSU81. - Yeast hypothetical proteins YAR014c (1 copy), YFR024c (1 copy), YHL002w (1 copy), YHR016c (1 copy), YJL020C (1 copy), YHR114w (2 copies) and the fission yeast homolog SpAC12C2.05c. - *Caenorhabditis elegans* hypothetical proteins F42H10.3. The profile developed to detect SH3 domains is based on a structural alignment consisting of 5 gap-free blocks and 4 linker regions totaling 62 match positions.

[1] Mayer B.J., Hamaguchi M., Hanafusa H. *Nature* 332:272-275(1988).[2] Musacchio A., Gibson T., Lehto V.P., Saraste M. *FEBS Lett.* 307:55-61(1992).[3] Pawson T., Schlessinger J. *Curr. Biol.* 3:434-442(1993).[4] Mayer B.J., Baltimore D. *Trends Cell Biol.* 3:8-13(1993).[5] Pawson T. *Nature* 373:573-580(1995).[6] Kuriyan J., Cowburn D. *Curr. Opin. Struct. Biol.* 3:828-837(1993).[7] Morton C.J., Campbell I.D. *Curr. Biol.* 4:615-617(1994).

590. Serine hydroxymethyltransferase pyridoxal-phosphate attachment site (SHMT)
Serine hydroxymethyltransferase (EC 2.1.2.1) (SHMT) [1] catalyzes the transfer of the hydroxymethyl group of serine to tetrahydrofolate to form 5,10-methylenetetrahydrofolate and glycine. In vertebrates, it exists in acytoplasmic and a mitochondrial form whereas only one form is found in prokaryotes. Serine hydroxymethyltransferase is a pyridoxal-phosphate containing enzyme. The pyridoxal-P group is attached to a lysine residue around which the sequence is highly conserved in all forms of the enzyme.

Consensus pattern: [DEH]-[LIVMFY]-x-[STMV]-[GST]-[ST](2)-H-K-[ST]-[LF]-x-G-[PAC]-[RQ]-[GSA]-[GA] [K is the pyridoxal-P attachment site]
[1] Usha R., Savithri H.S., Rao N.A. *Biochim. Biophys. Acta* 1204:75-83(1994).

591. SIS domain

SIS (Sugar ISomerase) domains are found in many phosphosugar isomerases and phosphosugar binding proteins.

[1] Teplyakov A, Obmolova G, Badet-Denisot MA, Badet B, Polikarpov I; Structure 1998;6:1047-1055.

5 592. (SKI) Shikimate kinase signature

Shikimate kinase (EC 2.7.1.71) catalyzes the fifth step in the biosynthesis from chorismate of the aromatic amino acids (the shikimate pathway) in bacteria (gene *aroK* or *aroL*), plants and in fungi (where it is part of a multifunctional enzyme which catalyzes five consecutive steps in this pathway). Shikimate kinase is a small protein of about 200 residues. A conserved region that contains a run of three glycines has been selected as a signature pattern.

Consensus pattern: [KR]-x(2)-E-x(3)-[LIVMF]-x(8,12)-[LIVMF](2)-[SA]-x-G(3)-x-[LIVMF]. Proteins belonging to this family also contain a copy of the ATP/GTP-binding motif 'A' (P-loop).

593. SNAP-25 family

SNAP-25 (synaptosome-associated protein 25 kDa) proteins are components of SNARE complexes. Members of this family contain a cluster of cysteine residues that can be palmitoylated for membrane attachment [2].

[1] Brennwald P, Kearns B, Champion K, Keranen S, Bankaitis V, Novick P; Cell 1994;79:245-258. [2] Risinger C, Blomqvist AG, Lundell I, Lambertsson A, Nassel D, Pieribone VA, Brodin L, Larhammar D; J Biol Chem 1993;268:24408-24414.

594. SNF2 and others N-terminal domain

This domain is found in proteins involved in a variety of processes including transcription regulation (e.g., SNF2, STH1, brahma, MOT1), DNA repair (e.g., ERCC6, RAD16, RAD5), DNA recombination (e.g., RAD54), and chromatin unwinding (e.g., ISWI) as well as a variety of other proteins with little functional information (e.g., Iodestar, ETL1).

595. Staphylococcal nuclease homologues (Snase)

Present in all three domains of cellular life. Four copies in the transcriptional coactivator p100. These, however, appear to lack the active site residues of Staphylococcal nuclease.

Positions 14 (Asp-21), 34 (Arg-35), 39 (Asp-40), 42 (Glu-43) and Comment: 110 (Arg-87) [SNase numbering in parentheses] are thought to be involved in substrate-binding and catalysis.

[1] Ponting CP; Protein Sci 1997;6:459-463. [2] Callebaut I, Mornon JP; Biochem J 1997;321:125-132.

596. SPRY domainA

SPRY Domain is named from SPla and the RYanodine Receptor. Domain of unknown function. Distant homologues are domains in Comment: butyrophilin/marenostrin/pyrin homologues.

[1] Ponting C, Schultz J, Bork P; Trends Biochem Sci 1997;22:193-194.

597. (SQS PSY) Squalene and phytoene synthases signatures

Two different polyisoprene synthases have been shown [1,2,3] to share a number of regions of sequence similarities: - Squalene synthase (EC 2.5.1.21) (farnesyl-diphosphate farnesyltransferase) (SQS), which catalyzes the conversion of two molecules of farnesyl diphosphate (FPP) into squalene. It is the first committed step in the cholesterol biosynthetic pathway. The reaction carried out by SQS is catalyzed in two separate steps: the first is a head-to-head condensation of the two molecules of FPP to form presqualene diphosphate; this intermediate is then rearranged in a NADP-dependent reduction, to form squalene. SQS is found in eukaryotes. In yeast it is encoded by the ERG9 gene, in mammals by the FDFT1 gene. SQS seems to be membrane-bound. - Phytoene synthase (EC 2.5.1.-) (PSY), which catalyzes the conversion of two molecules of geranylgeranyl diphosphate (GGPP) into phytoene. It is the second step in the biosynthesis of carotenoids from isopentenyl diphosphate. The reaction carried out by PSY is catalyzed in two separate steps: the first is a head-to-head condensation of the two molecules of GGPP to form prephytoene diphosphate; this intermediate is then rearranged to form phytoene. PSY is found in all organisms that

synthesize carotenoids: plants and photosynthetic bacteria as well as some non-photosynthetic bacteria and fungi. In bacteria PSY is encoded by the gene crtB. In plants PSY is localized in the chloroplast. As it can be seen from the description above, both SQS and PSY share a number of functional similarities which are also reflected at the level of their primary structure. In particular three well conserved regions are shared by SQS and PSY; they could be involved in substrate binding and/or the catalytic mechanism. Signature patterns have been developed for the second and third conserved regions; they are localized in the central part of these enzymes.

Consensus pattern: Y-[CSAM]-x(2)-[VSG]-A-[GSA]-[LIVAT]-[IV]-G-x(2)-[LMSC]- x(2)-[LIV]

Consensus pattern: [LIVM]-G-x(3)-Q-x(2,3)-N-[IF]-x-R-D-[LIVMFY]-x(2)-[DE]- x(4,7)-R-x-[FY]-x-P-

[1] Summers C., Karst F., Charles A.D. Gene 136:185-192(1993).[2] Robinson G.W., Tsay Y.H., Kienzle B.K., Smith-Monroy C.A., Bishop R.W. Mol. Cell. Biol. 13:2706-2727(1993).[3] Roemer S., Hugueney P., Bouvier F., Camara B., Kuntz M. Biochem. Biophys. Res. Commun. 196:1414-1421(1993).

598. SRP54-type proteins GTP-binding domain signature

The signal recognition particle (SRP) is an oligomeric complex that mediates targeting and insertion of the signal sequence of exported proteins into the membrane of the endoplasmic reticulum. SRP consists of a 7S RNA and six protein subunits. One of these subunits, the 54 Kd protein (SRP54), is a GTP-binding protein that interacts with the signal sequence when it emerges from the ribosome. The N-terminal 300 residues of SRP54 include the GTP-binding site (G-domain) and are evolutionary related to similar domains in other proteins which are listed below [1]. - Escherichia coli and Bacillus subtilis ffh protein (P48), a protein which seems to be the prokaryotic counterpart of SRP54. Ffh is associated with a 4.5S RNA in the prokaryotic SRP complex. - Signal recognition particle receptor alpha subunit (docking protein), an integral membrane GTP-binding protein which ensures, in conjunction with SRP, the correct targeting of nascent secretory proteins to the endoplasmic reticulum membrane. The G-domain is located at the C-terminal extremity of the protein. - Bacterial ftsY protein, a protein which is believed to play a similar role to that of the docking protein in eukaryotes. The G-domain is located at the C-terminal extremity of the protein. - The pilA protein from

Neisseria gonorrhoeae which seems to be the homolog of ftsY. - A protein from the archaeobacteria Sulfolobus solfataricus. This protein is also believed to be a docking protein. The G-domain is also at the C- terminus. - Bacterial flagellar biosynthesis protein flhF. The best conserved regions in those domains are the sequence motifs that are part of the GTP-binding site, but as those regions are not specific to these proteins, they were not used as a signature pattern. Instead, a conserved region located at the C-terminal end of the domain was selected.

Consensus pattern: P-[LIVM]-x-[FYL]-[LIVMAT]-[GS]-x-[GS]-[EQ]-x(4)-[LIVMF]
[1] Althoff S., Selinger D., Wise J.A. Nucleic Acids Res. 22:1933-1947(1994).

599. (STphosphatase) Serine/threonine specific protein phosphatases signature

Serine/threonine specific protein phosphatases (EC 3.1.3.16) (PP) [1,2,3] are enzymes that catalyze the removal of a phosphate group attached to a serine or evolutionary related. - Protein phosphatase-1 (PP1) is an enzyme of broad specificity. It is inhibited by two thermostable proteins, inhibitor-1 and -2. In mammals, there are two closely related isoforms of PP-1: PP-1alpha and PP-1beta, produced by alternative splicing of the same gene. In Emericella nidulans, PP-1 (gene bimG) plays an important role in mitosis control by reversing the action of the nimA kinase. In yeast, PP-1 (gene SIT4) is involved in dephosphorylating the large subunit of RNA polymerase II. - Protein phosphatase-2A (PP2A) is also an enzyme of broad specificity. PP2A is a trimeric enzyme that consist of a core composed of a catalytic subunit associated with a 65 Kd regulatory subunit and a third variable subunit. In mammals, there are two closely related isoforms of the catalytic subunit of PP2A: PP2A-alpha and PP2A-beta, encoded by separate genes. - Protein phosphatase-2B (PP2B or calcineurin), a calcium-dependent enzyme whose activity is stimulated by calmodulin. It is composed of two subunits: the catalytic A-subunit and the calcium-binding B-subunit. The specificity of PP2B is restricted. In addition to the above-mentioned enzymes, some additional serine/threonine specific protein phosphatases have been characterized and are listed below. - Mammalian phosphatase-X (PP-X), and Drosophila phosphatase-V (PP-V) which are closely related but yet distinct from PP2A. - Yeast phosphatase PPH3, which is similar to PP2A, but with different enzymatic properties. - Drosophila phosphatase-Y (PP-Y), and yeast phosphatases Z1 and Z2 (genes PPZ1 and PPZ2) which are closely related but yet distinct from PP1. - Drosophila retinal degeneration protein C (gene rdgC), a calcium-binding

phosphatase required to prevent light-induced retinal degeneration. - Phages Lambda and Phi-80 ORF-221 which have been shown to have phosphatase activity and are related to mammalian PP's. The best conserved regions in these proteins is a perfectly conserved pentapeptide that can be used as a signature pattern.

5 Consensus pattern: [LIVM]-R-G-N-H-E-

[1] Cohen P. Annu. Rev. Biochem. 58:453-508(1989).[2] Cohen P., Cohen P.T.W. J. Biol. Chem. 264:21435-21438(1989).[3] Cohen P.T.W., Brewis N.D., Hughes V., Mann D.J. FEBS Lett. 268:355-359(1990).

10

600. Translation initiation factor SUI1 signature

In budding yeast (*Saccharomyces cerevisiae*), SUI1 is a translation initiation factor that functions in concert with eIF-2 and the initiator tRNA-Met in directing the ribosome to the proper start site of translation [1]. SUI1 is a protein of 108 residues. Close homologs of SUI1 have been found [2] in mammals, insects and plants. SUI1 is also evolutionary related to hypothetical proteins from *Escherichia coli* (*yciH*), *Haemophilus influenzae* (HI1225) and *Methanococcus vannielii*. A conserved region in the C-terminal section has been selected as a signature pattern.

Consensus pattern: [LIVM]-[EQ]-[LIVM]-Q-G-[DEN]-[KHQ]-[KRV]

[1] Yoon H., Donahue T.F. Mol. Cell. Biol. 12:248-260(1992).[2] Fields C.A., Adams M.D. Biochem. Biophys. Res. Commun. 198:288-291(1994).

25

601. (S T dehydratase) Serine/threonine dehydratases pyridoxal-phosphate attachment site
Serine and threonine dehydratases [1,2] are functionally and structurally related pyridoxal-phosphate dependent enzymes: - L-serine dehydratase (EC 4.2.1.13) and D-serine

dehydratase (EC 4.2.1.14) catalyze the dehydration of L-serine (respectively D-serine) into ammonia and pyruvate. - Threonine dehydratase (EC 4.2.1.16) (TDH) catalyzes the dehydration of threonine into alpha-ketobutarate and ammonia. In *Escherichia coli* and

30

other microorganisms, two classes of TDH are known to exist. One is involved in the biosynthesis of isoleucine, the other in hydroxamino acid catabolism. Threonine synthase (EC 4.2.99.2) is also a pyridoxal-phosphate enzyme, it catalyzes the transformation of homoserine-phosphate into threonine. It has been shown [3] that threonine synthase is

distantly related to the serine/threonine dehydratases. In all these enzymes, the pyridoxal-phosphate group is attached to a lysine residue. The sequence around this residue is sufficiently conserved to allow the derivation of a pattern specific to serine/threonine dehydratases and threonine synthases.

- 5 Consensus pattern: [DESH]-x(4,5)-[STVG]-x-[AS]-[FYI]-K-[DLIFSA]-[RVMF]-[GA]-[LIVMGA] [The K is the pyridoxal-P attachment site]
 [1] Ogawa H., Gomi T., Konishi K., Date T., Naakashima H., Nose K., Matsuda Y., Peraino C., Pitot H.C., Fujioka M. J. Biol. Chem. 264:15818-15823(1989).[2] Datta P., Goss T.J., Omnaas J.R., Patil R.V. Proc. Natl. Acad. Sci. U.S.A. 84:393-397(1987).[3] Parsot C. EMBO J. 5:3013-3019(1986).[4] Grabowski R., Hofmeister A.E.M., Buckel W. Trends Biochem. Sci. 18:297-300(1993).

Cysteine synthase/cystathionine beta-synthase P-phosphate attachment site

Cysteine synthase (CSase) is the pyridoxal-phosphate dependent enzyme responsible [1] for the formation of cysteine from O-acetyl-serine and hydrogen sulfide with the concomitant release of acetic acid. In bacteria such as *Escherichia coli*, two forms of the enzyme are known (genes *cysK* and *cysM*). In plants there are also two forms, one located in the cytoplasm and the other in chloroplasts. Cystathionine beta-synthase [2] catalyzes the first irreversible step in homocysteine transulfuration; the conjugation of homocysteine and serine forming cystathionine. Like CSase it is a pyridoxal-phosphate dependent enzyme. The two types of enzymes are evolutionary related. The pyridoxal-phosphate group of CSases has been shown to be attached to a lysine residue which is located in the N-terminal section of these enzymes; the sequence around this residue is highly conserved and can be used as a signature pattern to detect this class of enzymes.

- 25 Consensus pattern: K-x-E-x(3)-[PA]-[STAGC]-x-S-[IVAP]-K-x-R-x-[STAG]-x(2)-[LIVM]
 [The 2nd K is the pyridoxal-P attachment site]
 [1] Saito K., Kurosawa M., Murakoshi I. FEBS Lett. 328:111-114(1993).[2] Swaroop M., Bradley K., Ohura T., Tahara T., Roper M.D., Rosenberg L.E., Kraus J.P. J. Biol. Chem. 267:11455-11461(1992).

30

602. S locus glycop

S-locus glycoprotein family. In Brassicaceae, self-incompatible plants have a self/non-self
 Comment: recognition system. This is sporophytically controlled by Comment: multiple
 alleles at a single locus (S). S-locus glycoproteins, Comment: as well as S-receptor kinases,
 are in linkage with the S-alleles [1]. Number of members: 128

- 5 [1] Evolutionary aspects of the S-related genes of the Brassica self-incompatibility system:
 synonymous and nonsynonymous base substitutions. Hinata K, Watanabe M, Yamakawa S,
 Satta Y, Isogai A; Genetics 1995;140:1099-1104. [2] Polymorphism of the S-locus
 glycoprotein gene (SLG) and the S-locus related gene (SLR1) in *Raphanus sativus* L. and
 self-incompatible ornamental plants in the Brassicaceae. Sakamoto K, Kusaba M, Nishio T;
 10 Mol Gen Genet 1998;258:397-403.

603. (sdh cyt) Succinate dehydrogenase cytochrome b subunit signatures

Succinate dehydrogenase (SDH) is a membrane-bound complex of two main components: a
 15 membrane-extrinsic component composed of an FAD-binding flavoprotein and an iron-sulfur
 protein, and a hydrophobic component composed of a cytochrome B and a membrane anchor
 protein. The cytochrome b component is a mono heme transmembrane protein [1,2,3]
 belonging to a family that groups: - Cytochrome b-556 from bacterial SDH (gene *sdhC*). -
 Cytochrome b560 from the mammalian mitochondrial SDH complex. - Cytochrome b560
 20 subunit encoded in the mitochondrial genome of some algae and in the plant *Marchantia*
polymorpha. - Cytochrome b from yeast mitochondrial SDH complex (gene *SDH3* or *CYB3*).
 - Protein cyt-1 from *Caenorhabditis*. These cytochromes are proteins of about 130 residues
 that comprise three transmembrane regions. There are two conserved histidines which may
 be involved in binding the heme group. Two signature patterns have been developed that
 25 include these histidine residues.

Consensus pattern: R-P-[LIVMT]-x(3)-[LIVM]-x(6)-[LIVMWPK]-x(4)-S-x(2)-H-R-x- [ST]
 [H could be a heme ligand]

Consensus pattern: H-x(3)-[GA]-[LIVMT]-R-[HF]-[LIVMF]-x-[FYWM]-D-x-[GVA] [H
 could be a heme ligand]

- 30 [1] Yu L., Wei Y.-Y., Usui S., Yu C.-A. J. Biol. Chem. 267:24508-24515(1992).[2]
 Abraham P.R., Mulder A., Van't Riet J., Raue H.A. Mol. Gen. Genet. 242:708-716(1994).[3]
 Leblanc C., Boyen C., Richard O., Bonnard G., Grienemberger J.M., Kloareg B. J. Mol. Biol.
 250:484-495(1995).

604. Sec1 family

[1] The Sec1 family: a novel family of proteins involved in synaptic transmission and general secretion. Halachmi N, Lev Z; J Neurochem 1996;66:889-897.

Number of members: 40

605. Protein secE/sec61-gamma signature

In bacteria, the secE protein plays a role in protein export; it is one of the components - with secY and secA - of the preprotein translocase. In eukaryotes, the evolutionary related protein sec61-gamma plays a role in protein translocation through the endoplasmic reticulum; it is part of a trimeric complex that also consist of sec61-alpha and beta [1]. Both secE and sec61-gamma are small proteins of about 60 to 90 amino acids that contain a single transmembrane region at their C-terminal extremity (Escherichia coli secE is an exception, in that it possess an extra N-terminal segment of 60 residues that contains two additional transmembrane domains). The sequence of secE/sec61-gamma is not extremely well conserved, however it is possible to derive a signature pattern centered on a conserved proline located 10 residues before the beginning of the transmembrane domain.

Consensus pattern: [LIVMFY]-x(2)-[DENQGA]-x(4)-[LIVMFTA]-x-[KRV]-x(2)-[KW]-P-x(3)-[SEQ]-x(7)-[LIVT]-[LIVGA]-[LIVFGAST]

[1] Hartmann E., Sommer T., Prehn S., Goerlich D., Jentsch S., Rapoport T.A. Nature 367:654-657(1994).

606. 11-S plant seed storage proteins signature

Plant seed storage proteins, whose principal function appears to be the major nitrogen source for the developing plant, can be classified, on the basis of their structure, into different families. 11-S are non-glycosylated proteins which form hexameric structures [1,2]. Each of the subunits in the hexamer is itself composed of an acidic and a basic chain derived from a single precursor and linked by a disulfide bond. This structure is shown in the following representation. +-----+ ||

xxxxxxxxxxCxxxxxxxxxxxxxxxxxxxxxxxxxxNGxCxxxxxxxxxxxxxxxxxxxxxxxxxx ***** <--

----Acidic-subunit-----><----Basic-subunit-----> <-----About-480-to-500-residues----->'C': conserved cysteine involved in a disulfide bond.'*': position of the pattern. Proteins that belong to the 11-S family are: pea and broad bean legumins, rape cruciferin, rice glutelins, cotton beta-globulins, soybean glycinins, pumpkin 11-S globulin, oat globulin, sunflower helianthinin G3, etc. The region that includes the conserved cleavage site between the acidic and basic subunits (Asn-Gly) and a proximal cysteine residue which is involved in the interchain disulfide bond have been used as a signature pattern for this family of proteins.

Consensus pattern: N-G-x-[DE](2)-x-[LIVMF]-C-[ST]-x(11,12)-[PAG]-D [C is involved in a disulfide bond

[1] Hayashi M., Mori H., Nishimura M., Akazawa T., Hara-Nishimura I. Eur. J. Biochem. 172:627-632(1988).[2] Shotwell M.A., Afonso C., Davies E., Chesnut R.S., Larkins B.A. Plant Physiol. 87:698-704(1988).

607. 7S seed storage protein

7S globulin is one of the main storage proteins of most angiosperms and gymnosperms. The 7S storage proteins are homotrimers.

Number of members: 67

[1] The three-dimensional structure of canavalin from jack bean (*Canavalia ensiformis*). Ko TP, Ng JD, McPherson A; Plant Physiol 1993;101:729-744.

608. Aspartate-semialdehyde dehydrogenase signature

Aspartate-semialdehyde dehydrogenase (ASD) catalyzes the second step in the common biosynthetic pathway leading from Asp to diaminopimelate and Lys, to Met, and to Thr; the NADP-dependent reductive dephosphorylation of L-aspartyl phosphate to L-aspartate-semialdehyde. In bacteria and fungi, ASD is a protein of about 40 Kd (340 to 370 residues) whose sequence is not extremely well conserved [1]. A conserved cysteine residue has been implicated as important for the catalytic activity [2]. The region of conservation around the active site residue is too small to be used as signature pattern. Another more conserved region, located in the last third of the sequence, and which contains both a conserved cysteine as well as an histidine has been used instead.

Consensus pattern: [LIVM]-[SADN]-x(2)-C-x-R-[LIVM]-x(4)-[GSC]-H-[STA

[1] Baril C., Richaud C., Fourni E., Baranton G., Saint Girons I. J. Gen. Microbiol. 138:47-53(1992).[2] Karsten W.E., Viola R.E. Biochim. Biophys. Acta 1121:234-238(1992).

5 N-acetyl-gamma-glutamyl-phosphate reductase active site

N-acetyl-gamma-glutamyl-phosphate reductase (EC 1.2.1.38) (AGPR) [1,2] is the enzyme that catalyzes the third step in the biosynthesis of arginine from glutamate, the NADP-dependent reduction of N-acetyl-5-glutamyl phosphate into N-acetylglutamate 5-semialdehyde. In bacteria it is a monofunctional protein of 35 to 38 Kd (gene argC) while in
10 fungi it is part of a bifunctional mitochondrial enzyme (gene ARG5,6, arg11 or arg-6) which contains a N-terminal acetylglutamate kinase (EC 2.7.2.8) domain and a C-terminal AGPR domain. In the Escherichia coli enzyme, a cysteine has been shown to be implicated in the catalytic activity, the region around this residue is well conserved and can be used as a signature pattern.

15 Consensus pattern: [LIVM]-[GSA]-x-P-G-C-[FY]-[AVP]-T-[GA]-x(3)-[GTAC]-[LIVM]- x-P [C is the active site residue]

[1] Ludovice M., Martin J.F., Carrachas P., Liras P. J. Bacteriol. 174:4606-4613(1992).[2] Gessert S.F., Kim J.H., Nargang F.E., Weiss R.L. J. Biol. Chem. 269:8189-8203(1994).

20 609. Sialyltransferase family,

Number of members: 18

25 610. SpoU rRNA Methylase family

This family of proteins probably use S-AdoMet. Number of members: 58

[1] SpoU protein of Escherichia coli belongs to a new family of putative rRNA methylases. Koonin EV, Rudd KE; Nucleic Acids Res 1993;21:5519-5519. [2] The spoU gene of escherichia coli , the fourth gene of the spoT operon, is essential for tRNA (Gm18) 2 '
30 methyltransferase activity. Persson BC, Jager G, Gustafsson C; Nucleic Acids Res 1997;25:4093-4097.

611. Stathmin family signatures

Stathmin [1] (from the Greek 'stathmos' which means relay), is an ubiquitous intracellular protein, present in a variety of phosphorylated forms and which serves as a relay for diverse second messenger pathways. Its expression and phosphorylation are regulated throughout development and in response to extracellular signals regulating cell proliferation, differentiation and function. Stathmin is a highly conserved protein of 149 amino acid residues. Structurally, it consists of an N-terminal domain of about 45 residues followed by a 78 residue alpha-helical domain consisting of a heptad repeat coiled coil structure and a C-terminal domain of 25 residues. Protein SCG10 is a neuron-specific, membrane-associated protein that accumulates in the growth cones of developing neurons. It is highly similar in its sequence to stathmin, but differs in that it contains an additional N-terminal hydrophobic segment of 32 residues which is probably responsible for its interaction with membranes. *Xenopus* protein XB3 is also evolutionary related to stathmin and also contains an additional N-terminal hydrophobic domain [2]. A conserved decapeptide which ends with the first three residues of the coiled coil domain and a second pattern that corresponds to part of the central region of the coiled coil have been selected as signatures for proteins of the stathmin family. Consensus pattern: P-[KRQ]-[KR](2)-[DE]-x-S-L-[EG]-E- Consensus pattern: A-E-K-R-E-H-E-[KR]-E- [1] Sobel A. Trends Biochem. Sci. 16:301-305(1991).[2] Maucuer A., Moreau J., Mechali M., Sobel A. J. Biol. Chem. 268:16420-16429(1993).

612. SUA5/yciO/yrdC family signature. The following uncharacterized proteins have been shown [1] to share regions of similarities: - Yeast protein SUA5. - *Escherichia coli* hypothetical protein yciO and HI1198, the corresponding *Haemophilus influenzae* protein. - *Escherichia coli* hypothetical protein yrdC and HI0656, the corresponding *Haemophilus influenzae* protein. - *Bacillus subtilis* hypothetical protein ywlC. - *Mycobacterium leprae* hypothetical protein in rfe-hemK intergenic region. - *Methanococcus jannaschii* hypothetical protein MJ0062. These are proteins of from 20 to 46 Kd which contain a number of conserved regions in their N-terminal section. They can be picked up in the database by the following pattern.

511

Consensus pattern: [LIVMTA](3)-[LIVMFYC]-[PG]-T-[DE]-[STA]-x-[FY]-[GA]- [LIVM]-
[GS]-

[1] Bairoch A., Rudd K.E., Robison K. Unpublished observations (1995).

5

613. Sucrose synthase

Sucrose synthases catalyse the synthesis of sucrose from UDP-glucose and fructose. This family includes the bulk of the sucrose synthase protein. However the carboxyl terminal region of the sucrose synthases belongs to the glycosyl transferase family Glycos transf 1.

10

614. Sulfotransferase proteins

Number of members: 59

5

615. Synaptophysin / synaptoporin signature

Synaptophysin and synaptoporin [1] are structurally related proteins, found in the membrane of synaptic vesicles, which may function as ionic or solute channels. These two glycoproteins seem to span the membrane four times. Both their N- and C-termini sequences seem to be cytoplasmically located. As a signature pattern for this family of proteins, a highly conserved region located in the beginning of the first intravesicular loop just after the first transmembrane domain has been selected. This region contains a cysteine residue that may be involved in a disulfide bond.

20

Consensus pattern: L-S-V-[DE]-C-x-N-K-T [C may be involved in a disulfide bond

[1] Knaus P., Marqueze-Pouey B., Scherer H., Betz H. Neuron 5:453-462(1990).

25

616. Syndecans signature

Syndecans [1,2] (from the greek syndein; to bind together) are a family of transmembrane heparan sulfate proteoglycans which are implicated in the binding of extracellular matrix components and growth factors. Syndecans bind a variety of molecules via their heparan sulfate chains and can act as receptors or as co-receptors. Structurally, these proteins consist

30

of four separate domains: a) A signal sequence; b) An extracellular domain (ectodomain) of variable length and whose sequence is not evolutionary conserved in the various forms of syndecans. The ectodomain contains the sites of attachment of the heparan sulfate glycosaminoglycan side chains; c) A transmembrane region; d) A highly conserved cytoplasmic domain of about 30 to 35 residues which could interact with cytoskeletal proteins. The proteins known to belong to this family are: - Syndecan 1. - Syndecan 2 or fibroglycan. - Syndecan 3 or neuroglycan or N-syndecan. - Syndecan 4 or amphiglycan or ryudocan. - Drosophila syndecan. - Caenorhabditis elegans probable syndecan (F57C7.3). The signature pattern that has been developed for syndecans starts with the last residue of the transmembrane region and includes the first 10 residues of the cytoplasmic domain. This region, which contains four basic residues, could act as a stop transfer site.

Consensus pattern: [FY]-R-[IM]-[KR]-K(2)-D-E-G-S-Y

[1] Bernfield M., Kokenyesi R., Kato M., Hinkes M.T., Spring J., Gallo R.L., Lose E.J. Annu. Rev. Cell Biol. 8:365-393(1992).[2] David G. FASEB J. 7:1023-1030(1993).

617. Syntaxin / epimorphin family signature

The following proteins have been shown to be evolutionary related [1,2,3]: - Epimorphin (or syntaxin 2), a mammalian mesenchymal protein which plays an essential role in epithelial morphogenesis. - Syntaxin 1A (also known as antigen HPC-1) and syntaxin 1B which are synaptic proteins which may be involved in docking of synaptic vesicles at presynaptic active zones. - Syntaxin 3. - Syntaxin 4, which is potentially involved in docking of synaptic vesicles at presynaptic active zones. - Syntaxin 5, which mediates endoplasmic reticulum to golgi transport. - Syntaxin 6, which is involved in intracellular vesicle trafficking. - Syntaxin 7. - Yeast PEP12 (or VPS6) which is required for the transport of proteases to the vacuole. - Yeast SED5 which is required for the fusion of transport vesicles with the Golgi complex. - Yeast SSO1 and SSO2 which are required for vesicle fusion with the plasma membrane. - Yeast VAM3, which is required for vacuolar assembly. - Arabidopsis thaliana protein KNOLLE which may be involved in cytokinesis. - Caenorhabditis elegans hypothetical proteins F35C8.4, F48F7.2, F55A11.2 and T01B11.3. The above proteins share the following characteristics: a size ranging from 30 Kd to 40 Kd; a C-terminal extremity which is highly hydrophobic and is probably involved in anchoring the protein to the membrane; a central,

well conserved region, which seems to be in a coiled-coil conformation. The pattern specific for this family is based on the most conserved region of the coiled coil domain.

Consensus pattern: [RQ]-x(3)-[LIVMA]-x(2)-[LIVM]-[ESH]-x(2)-[LIVMT]-x-[DEVMT]-
[LIVM]-x(2)-[LIVM]-[FS]-x(2)-[LIVM]-x(3)-[LIVT]-x(2)-Q- [GADEQ]-x(2)-[LIVM]-
5 [DNQT]-x-[LIVMF]-[DESV]-x(2)-[LIVM]

[1] Bennett M.K., Garcia-Arraras J.E., Elferink L.A., Peterson K., Fleming A.M., Hazuka C.D., Scheller R.H. Cell 74:863-873(1993). [2] Spring J., Kato M., Bernfield M. Trends Biochem. Sci. 18:124-125(1993).[3] Pelham H.R.B. Cell 73:425-426(1993).

10

618. Sm protein

The U1, U2, U4/U6, and U5 small nuclear ribonucleoprotein particles (snRNPs) involved in pre-mRNA splicing contain seven Sm proteins (B/B', D1, D2, D3, E, F and G) in common, which
15 assemble around the Sm site present in four of the major spliceosomal small nuclear RNAs. These proteins contain a common sequence motif in two segments, Sm1 and Sm2, separated by a short variable linker.

20

[1] Hermann H, Fabrizio P, Raker VA, Foulaki K, Hornig H, Brahms H, Luhrmann R EMBO J 1995;14:2076-2088. [2] Kambach C, Walke S, Young R, Avis JM, de la Fortelle E, Raker VA, Luhrmann R, Li J, Nagai K; Cell 1999;96:375-387.

25

619. Skp1 family

[1] Stebbins CE, Kaelin WG Jr, Pavletich NP; Science 1999;284:455-461.

30

620. Protein secY signatures

The eubacterial secY protein [1] plays an important role in protein export. It interacts with the signal sequences of secretory proteins as well as with two other components of the protein translocation system: secA and secE. SecY is an integral plasma membrane protein of 419 to

492 amino acid residues that apparently contains ten transmembrane segments. Such a structure probably confers to secY a 'translocator' function, providing a channel for periplasmic and outer-membrane precursor proteins. Homologs of secY are found in archaeobacteria [2]. SecY is also encoded in the chloroplast genome of some algae [3] where it could be involved in a prokaryotic-like protein export system across the two membranes of the chloroplast endoplasmic reticulum (CER) which is present in chromophyte and cryptophyte algae. Two signature patterns have been developed for secY proteins. The first corresponds to the second transmembrane region, which is the most conserved section of these proteins. The second spans the C-terminal part of the fourth transmembrane region, a short intracellular loop, and the N-terminal part of the fifth transmembrane region.

Consensus pattern: [GST]-[LIVMF](2)-x-[LIVM]-G-[LIVM]-x-P-[LIVMFY](2)-x-[AS]-[GSTQ]-[LIVMFAT](3)-Q-[LIVMFA](2)

Consensus pattern: [LIVMFYW](2)-x-[DE]-x-[LIVMF]-[STN]-x(2)-G-[LIVMF]-[GST]-[NST]-G-x-[GST]-[LIVMF](3)

[1] Ito K. Mol. Microbiol. 6:2423-2428(1992).[2] Auer J., Spicker G., Boeck A. Biochimie 73:683-688(1991).[3] Douglas S.E. FEBS Lett. 298:93-96(1992).

621. (Seed protein) Small hydrophilic plant seed proteins signature. The following small hydrophilic plant seed proteins are structurally related: - Arabidopsis thaliana proteins GEA1 and GEA6. - Cotton late embryogenesis abundant (LEA) protein D-19. - Carrot EMB-1 protein. - Barley LEA proteins B19.1A, B19.1B, B19.3 and B19.4. - Maize late embryogenesis abundant protein Emb564. - Radish late seed maturation protein p8B6. - Rice embryonic abundant protein Emp1. - Sunflower 10 Kd late embryogenesis abundant protein (DS10). - Wheat Em proteins. These proteins contain from 83 to 153 amino acid residues and may play a role [1,2] in equipping the seed for survival, maintaining a minimal level of hydration in the dry organism and preventing the denaturation of cytoplasmic components. They may also play a role during imbibition by controlling water uptake. As a signature pattern, the best conserved region in the sequence of these proteins has been developed, it is a glycine-rich nonapeptide located in the N-terminal section.-

Consensus pattern: G-[EQ]-T-V-V-P-G-G-T-

[1] Dure L. III, Crouch M., Harada J., Ho T.-H. D., Mundy J., Quatrano R., Thomas T., Sung Z.R. *Plant Mol. Biol.* 12:475-486(1989).[2] Gaubier P., Raynal M., Hull G., Huestis G.M., Grellet F., Arenas C., Pages M., Delseny M. *Mol. Gen. Genet.* 238:409-418(1993).

5

622. Serine carboxypeptidases, active sites

All known carboxypeptidases are either metallo carboxypeptidases or serinecarboxypeptidases. The catalytic activity of the serine carboxypeptidases, like that of the trypsin family serine proteases, is provided by a charge relay system involving an aspartic acid residue hydrogen-bonded to a histidine, which is itself hydrogen-bonded to a serine [1].

10

Proteins known to be serine carboxypeptidases are: - Barley and wheat serine carboxypeptidases I, II, and III [2]. - Yeast carboxypeptidase Y (YSCY) (gene PRC1), a vacuolar protease involved in degrading small peptides. - Yeast KEX1 protease, involved in killer toxin and alpha-factor precursor processing. - Fission yeast *sxa2*, a probable carboxypeptidase involved in degrading or processing mating pheromones [3]. - *Penicillium janthinellum* carboxypeptidase S1 [4]. - *Aspergillus niger* carboxypeptidase pepF. -

15

*Aspergillus sato*i carboxypeptidase cpdS. - Vertebrate protective protein / cathepsin A [5], a lysosomal protein which is not only a carboxypeptidase but also essential for the activity of both beta-galactosidase and neuraminidase. - Mosquito vitellogenic carboxypeptidase (VCP) [6]. - *Naegleria fowleri* virulence-related protein Nf314 [7]. - Yeast hypothetical protein YBR139w. - *Caenorhabditis elegans* hypothetical proteins C08H9.1, F13D12.6, F32A5.3, F41C3.5 and K10B2.2. This family also includes: - Sorghum (s)-hydroxymandelonitrile lyase (hydroxynitrile lyase) (HNL) [8], an enzyme involved in plant cyanogenesis. The sequences surrounding the active site serine and histidine residues are highly conserved in all these

20

serine carboxypeptidases.

25

Consensus pattern: [LIVM]-x-[GTA]-E-S-Y-[AG]-[GS] [S is the active site residue]

Consensus pattern: [LIVF]-x(2)-[LIVSTA]-x-[IVPST]-x-[GSDNQL]-[SAGV]-[SG]-H-x-[IVAQ]-P-x(3)-[PSA] [H is the active site residue]

30

[1] Liao D.I., Remington S.J. *J. Biol. Chem.* 265:6528-6531(1990).[2] Sorensen S.B., Svendsen I., Breddam K. *Carlsberg Res. Commun.* 54:193-202(1989).[3] Imai Y., Yamamoto M. *Mol. Cell. Biol.* 12:1827-1834(1992).[4] Svendsen I., Hofmann T., Endrizzi J., Remington J., Breddam K. *FEBS Lett.* 333:39-43(1993).[5] Galjart N.J., Morreau H., Willemsen R., Gillemans N., Bonten E.J., d'Azzo A. *J. Biol. Chem.* 266:14754-14762(1991).[

6] Cho W.L., Deitsch K.W., Raikhel A.S. Proc. Natl. Acad. Sci. U.S.A. 88:10821-10824(1991).[7] Hu W.N., Kopachik W., Band R.N. Infect. Immun. 60:2418-2424(1992).[8] Wajant H., Mundry K.W., Pfitzenmaier K. Plant Mol. Biol. 26:735-746(1994).[9] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).[E1]

5

623. Serpins signature. Serpins (SERine Proteinase INhibitors) [1,2,3,4] are a group of structurally related proteins. They are high molecular weight (400 to 500 amino acids), extracellular, irreversible serine protease inhibitors with a well defined structural-functional characteristic: a reactive region that acts as a 'bait' for an appropriate serine protease. This region is found in the C-terminal part of these proteins. Proteins which are known to belong to the serpin family are listed below (references are only provided for recently determined sequences): - Alpha-1 protease inhibitor (alpha-1-antitrypsin, contrapsin). - Alpha-1-antichymotrypsin, - Antithrombin III. - Alpha-2-antiplasmin. - Heparin cofactor II. - Complement C1 inhibitor. - Plasminogen activator inhibitors 1 (PAI-1) and 2 (PAI-2). - Glia derived nexin (GDN) (Protease nexin I). - Protein C inhibitor. - Rat hepatocytes SPI-1, SPI-2 and SPI-3 inhibitors. - Human squamous cell carcinoma antigen (SCCA) which may act in the modulation of the host immune response against tumor cells. - A lepidopteran protease inhibitor. - Leukocyte elastase inhibitor which, in contrast to other serpins, is an intracellular protein. - Neuroserpin [5], a neuronal inhibitor of plasminogen activators and plasmin. - Cowpox virus crmA [6], an inhibitor of the thiol protease interleukin-1B converting enzyme (ICE). CrmA is the only serpin known to inhibit a non-serine proteinase. - Some orthopoxviruses probable protease inhibitors, which may be involved in the regulation of the blood clotting cascade and/or of the complement cascade in the mammalian host. On the basis of strong sequence similarities, a number of proteins with no known inhibitory activity are said to belong to this family: - Birds ovalbumin and the related genes X and Y proteins. - Angiotensinogen; the precursor of the angiotensin active peptide. - Barley protein Z; the major endosperm albumin. - Corticosteroid binding globulin (CBG). - Thyroxine-binding globulin (TBG). - Sheep uterine milk protein (UTMP) and pig uteroferrin-associated protein (UFAP). - Hsp47, an endoplasmic reticulum heat-shock protein that binds strongly to collagen and could act as a chaperone in the collagen biosynthetic pathway [7]. - Maspin, which seems to function as a tumor suppressor [5]. - Pigment epithelium-derived factor precursor (PEDF), a protein with a strong neutrophilic activity [8]. -

10

15

20

25

30

Ep45, an estrogen-regulated protein from *Xenopus* [9]. A signature pattern has been developed for this family of proteins, centered on a well conserved Pro-Phe sequence which is found ten to fifteen residues on the C-terminal side of the reactive bond

- 5 Consensus pattern: [LIVMFY]-x-[LIVMFYAC]-[DNQ]-[RKHQS]-[PST]-F-[LIVMFY]-
[LIVMFYC]-x-[LIVMFAH]-

- 10 [1] Carrell R., Travis J. Trends Biochem. Sci. 10:20-24(1985).[2] Carrell R., Pemberton
P.A., Boswell D.R. Cold Spring Harbor Symp. Quant. Biol. 52:527-535(1987).[3] Huber R.,
Carrell R.W. Biochemistry 28:8951-8966(1989).[4] Remold-O'Donneel E. FEBS Lett.
315:105-108(1993).[5] Osterwalder T., Contartese J., Stoeckli E.T., Kuhn T.B., Sonderegger
P. EMBO J. 15:2944-2953(1996).[6] Komiyama T., Ray C.A., Pickup D.J., Howard A.D.,
Thornberry N.A., Peterson E.P., Salvesen G. J. Biol. Chem. 269:19331-19337(1994).[7]
Clarke E., Sandwal B.D. Biochim. Biophys. Acta 1129:246-248(1992).[8] Zou Z.,
15 Anisowicz A., Neveu M., Rafidi K., Sheng S., Sager R., Hendrix M.J., Seftor E., Thor A.
Science 263:526-529(1994).[9] Steele F.R., Chader G.J., Johnson L.V., Tombran-Tink J.
Proc. Natl. Acad. Sci. U.S.A. 90:1526-1530(1993).[10] Holland L.J., Suksang C., Wall A.A.,
Roberts L.R., Moser D.R., Bhattacharya A. J. Biol. Chem. 267:7053-7059(1992).

624. Sigma-54 interaction domain signatures and profile

Some bacterial regulatory proteins activate the expression of genes from promoters recognized by core RNA polymerase associated with the alternative sigma-54 factor. These have a conserved domain of about 230 residues involved in the ATP-dependent [1,2]

- 25 interaction with sigma-54. This domain has been found in the proteins listed below: - acoR
from *Alcaligenes eutrophus*, an activator of the acetoin catabolism operon acoXABC. - algB
from *Pseudomonas aeruginosa*, an activator of alginate biosynthetic gene algD. - dctD from
Rhizobium, an activator of dctA, the C4-dicarboxylate transport protein. - dhaR from
Citrobacter freundii, a regulator of the dha operon for glycerol utilization. - fh1A from
30 *Escherichia coli*, an activator of the formate dehydrogenase H and hydrogenase III structural
genes. - flbD from *Caulobacter crescentus*, an activator of flagellar genes. - hoxA from
Alcaligenes eutrophus, an activator of the hydrogenase operon. - hrpS from *Pseudomonas*
syringae, an activator of hprD as well as other hrp loci involved in plant pathogenicity. -

hupR1 from *Rhodobacter capsulatus*, an activator of the [NiFe] hydrogenase genes hupSL. - hydG from *Escherichia coli* and *Salmonella typhimurium*, an activator of the hydrogenase activity. - levR from *Bacillus subtilis*, which regulates the expression of the levanase operon (levDEFG and sacC). - nifA (as well as anfA and vnfA) from various bacteria, an activator of the nif nitrogen-fixing operon. - ntrC, from various bacteria, an activator of nitrogen assimilatory genes such as that for glutamine synthetase (glnA) or of the nif operon. - pgdA from *Salmonella typhimurium*, the activator of the inducible phospho- glycerate transport system. - pilR from *Pseudomonas aeruginosa*, an activator of pilin gene transcription. - rocR from *Bacillus subtilis*, an activator of genes for arginine utilization - tyrR from *Escherichia coli*, involved in the transcriptional regulation of aromatic amino-acid biosynthesis and transport. - wtsA, from *Erwinia stewartii*, an activator of plant pathogenicity gene wtsB. - xylR from *Pseudomonas putida*, the activator of the tol plasmid xylene catabolism operon xylCAB and of xylS. - *Escherichia coli* hypothetical protein yfhA. - *Escherichia coli* hypothetical protein yhgB. About half of these proteins (algB, dcdT, flbD, hoxA, hupR1, hydG, ntrC, pgdA and pilR) belong to signal transduction two-component systems [3] and possess a domain that can be phosphorylated by a sensor-kinase protein in their N- terminal section. Almost all of these proteins possess a helix-turn-helix DNA-binding domain in their C-terminal section. The domain which interacts with the sigma-54 factor has an ATPase activity. This may be required to promote a conformational change necessary for the interaction [4]. The domain contains an atypical ATP-binding motif A (P-loop) as well as a form of motif B. The two ATP-binding motifs are located in the N-terminal section of the domain; signature patterns have been developed for both motifs. Other regions of the domain are also conserved. One of them, located in the C-terminal section, has been selected as a third signature pattern.

Consensus pattern: [LIVMFY](3)-x-G-[DEQ]-[STE]-G-[STAV]-G-K-x(2)-[LIVMFY]

Consensus pattern: [GS]-x-[LIVMF]-x(2)-A-[DNEQASH]-[GNEK]-G-[STIM]-[LIVMFY](3)-[DE]-[EK]-[LIVM]

Consensus pattern: [FYW]-P-[GS]-N-[LIVM]-R-[EQ]-L-x-[NHAT]

[1] Morrett E., Segovia L. J. Bacteriol. 175:6067-6074(1993).[2] Austin S., Kundrot C.,

Dixon R. Nucleic Acids Res. 19:2281-2287(1991).[3] Albright L.M., Huala E., Ausubel F.M. Annu. Rev. Genet. 23:311-336(1989).[4] Austin S., Dixon R. EMBO J. 11:2219-2228(1992).

625. Sigma-70 factors family signatures

Sigma factors [1] are bacterial transcription initiation factors that promote the attachment of the core RNA polymerase to specific initiation sites and are then released. They alter the specificity of promoter recognition. Most bacteria express a multiplicity of sigma factors. Two of these factors, sigma-70 (gene *rpoD*), generally known as the major or primary sigma factor, and sigma-54 (gene *rpoN* or *ntrA*) direct the transcription of a wide variety of genes. The other sigma factors, known as alternative sigma factors, are required for the transcription of specific subsets of genes. With regard to sequence similarity, sigma factors can be grouped into two classes: the sigma-54 and sigma-70 families. The sigma-70 family includes, in addition to the primary sigma factor, a wide variety of sigma factors, some of which are listed below: - *Bacillus* sigma factors involved in the control of sporulation-specific genes: sigma-E (sigE or *spoIIGB*), sigma-F (sigF or *spoIIAC*), sigma-G (sigG or *spoIIIG*), sigma-H (sigH or *spo0C*) and sigma-K (sigK or *spoIVCB/spoIIIC*). - *Escherichia coli* and related bacteria sigma-32 (gene *rpoH* or *htpR*) involved in the expression of heat shock genes. - *Escherichia coli* and related bacteria sigma-27 (gene *fliA*) involved in the expression of the flagellin gene. - *Escherichia coli* sigma-S (gene *rpoS* or *katF*) which seems to be involved in the expression of genes required for protection against external stresses. - *Myxococcus xanthus* sigma-B (sigB) which is essential for the late-stage differentiation of that bacteria. Alignments of the sigma-70 family permit the identification of four regions of high conservation [2,3]. Each of these four regions can in turn be subdivided into a number of sub-regions. Signature patterns based on the two best-conserved sub-regions have been developed. The first pattern corresponds to sub-region 2.2; the exact function of this sub-region is not known although it could be involved in the binding of the sigma factor to the core RNA polymerase. The second pattern corresponds to sub-region 4.2 which seems to harbor a DNA-binding 'helix-turn-helix' motif involved in binding the conserved -35 region of promoters recognized by the major sigma factors. The second pattern starts one residue before the N-terminal extremity of the HTH region and ends six residues after its C-terminal extremity.

Consensus pattern: [DE]-[LIVMF](2)-[HEQS]-x-G-x-[LIVMFA]-G-L-[LIVMFYE]-x-[GSAM]-[LIVMAP]

Consensus pattern: [STN]-x(2)-[DEQ]-[LIVM]-[GAS]-x(4)-[LIVMF]-[PSTG]-x(3)-[LIVMA]-x-[NQR]-[LIVMA]-[EQH]-x(3)-[LIVMFW]-x(2)-[LIVM]

[1] Helmann J.D., Chamberlin M.J. *Annu. Rev. Biochem.* 57:839-872(1988).[2] Gribskov M., Burgess R.R. *Nucleic Acids Res.* 14:6745-6763(1986).[3] Lonetto M.A., Gribskov M., Gross C.A. *J. Bacteriol.* 174:3843-3849(1992).[4] Lonetto M.A., Brown K.L., Rudd K.E., Buttner M.J. *Proc. Natl. Acad. Sci. U.S.A.* 91:7573-7577(1994).

5

626. Signal carboxyl-terminal domain. 430 members.

10 627. Signal peptidases I signatures

Signal peptidases (SPases) [1] (also known as leader peptidases) remove the signal peptides from secretory proteins. In prokaryotes three types of SPases are known: type I (gene *lepB*) which is responsible for the processing of the majority of exported pre-proteins; type II (gene *lsp*) which only process lipoproteins, and a third type involved in the processing of pili subunits. SPase I is an integral membrane protein that is anchored in the cytoplasmic membrane by one (in *B. subtilis*) or two (in *E. coli*) N-terminal transmembrane domains with the main part of the protein protruding in the periplasmic space. Two residues have been shown [2,3] to be essential for the catalytic activity of SPase I: a serine and an lysine. SPase I is evolutionary related to the yeast mitochondrial inner membrane protease subunit 1 and 2 (genes *IMP1* and *IMP2*) which catalyze the removal of signal peptides required for the targeting of proteins from the mitochondrial matrix, across the inner membrane, into the inter-membrane space [4]. In eukaryotes the removal of signal peptides is effected by an oligomeric enzymatic complex composed of at least five subunits: the signal peptidase complex (SPC). The SPC is located in the endoplasmic reticulum membrane. Two

25 components of mammalian SPC, the 18 Kd (SPC18) and the 21 Kd (SPC21) subunits as well as the yeast SEC11 subunit have been shown [5] to share regions of sequence similarity with prokaryotic SPases I and yeast *IMP1/IMP2*. Three signature patterns for these proteins have been developed. The first signature contains the putative active site serine, the second signature contains the putative active site lysine which is not conserved in the SPC subunits,

30 and the third signature corresponds to a conserved region of unknown biological significance which is located in the C-terminal section of all these proteins.

Consensus pattern: [GS]-x-S-M-x-[PS]-[AT]-[LF] [S is an active site residue]

Consensus pattern: K-R-[LIVMSTA](2)-G-x-[PG]-G-[DE]-x-[LIVM]-x-[LIVMFY] [K is an active site residue]

Consensus pattern: [LIVMFYW](2)-x(2)-G-D-[NH]-x(3)-[SND]-x(2)-[SG]

[1] Dalbey R.E., von Heijne G. Trends Biochem. Sci. 17:474-478(1992).[2] Sung M., Dalbey R.E. J. Biol. Chem. 267:13154-13159(1992).[3] Black M.T. J. Bacteriol. 175:4957-4961(1993).[4] Nunnari J., Fox T.D., Walter P. Science 262:1997-2004(1993).[5] van Dijk J.M., de Jong A., Vehmaanpera J., Venema G., Bron S. EMBO J. 11:2819-2828(1992).[6] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).[E1]

10

628. (sodcu) Copper/Zinc superoxide dismutase signatures

Copper/Zinc superoxide dismutase (SODC) [1] is one of the three forms of an enzyme that catalyzes the dismutation of superoxide radicals. SODC binds one atom each of zinc and copper. Various forms of SODC are known: acytoplasmic form in eukaryotes, an additional chloroplast form in plants, an extracellular form in some eukaryotes, and a periplasmic form in prokaryotes. The metal binding sites are conserved in all the known SODC sequences [2]. Two signature patterns have been derived for this family of enzymes: the first one contains two histidine residues that bind the copper atom; the second one is located in the C-terminal section of SODC and contains a cysteine which is involved in a disulfide bond.

Consensus pattern: [GA]-[IMFAT]-H-[LIVF]-H-x(2)-[GP]-[SDG]-x-[STAGDE] [The two H's are copper ligands]

Consensus pattern: G-[GN]-[SGA]-G-x-R-x-[SGA]-C-x(2)-[IV] [C is involved in a disulfide bond]

[1] Bannister J.V., Bannister W.H., Rotilio G. CRC Crit. Rev. Biochem. 22:111-154(1987).[

2] Smith M.W., Doolittle R.F. J. Mol. Evol. 34:175-184(1992).

629. (sodfe) Manganese and iron superoxide dismutases signature

Manganese superoxide dismutase (SODM) [1] is one of the three forms of an enzyme that catalyzes the dismutation of superoxide radicals. The four ligands of the manganese atom are conserved in all the known SODM sequences. These metal ligands are also conserved in the related iron form of superoxide dismutases [2,3]. A short conserved region which includes two of the four ligands: an aspartate and a histidine has been selected as a signature.

522

Consensus pattern: D-x-W-E-H-[STA]-[FY](2) [D and H are manganese/iron ligands]

[1] Bannister J.V., Bannister W.H., Rotilio G. CRC Crit. Rev. Biochem. 22:111-154(1987).[

2] Parker M.W., Blake C.C.F. FEBS Lett. 229:377-382(1988).[3] Smith M.W., Doolittle

R.F. J. Mol. Evol. 34:175-184(1992).

5

630. Spectrin repeat

Spectrin repeats are found in several proteins involved in cytoskeletal structure. These include spectrin, alpha-actinin

10 and dystrophin. The sequence repeat used in this family is taken from the structural repeat in reference [2]. The spectrin repeat forms a three helix bundle. The second helix is interrupted by proline in some sequences.

Number of members: 898

[1] Actin-binding proteins. 1: Spectrin super family. Hartwig JH; Protein Profile 1995;2:732-732. [2] Crystal structure of the repetitive segments of spectrin. Yan Y, Winograd E, Viel A, Cronin T, Harrison SC, Branton D; Science 1993;262:2027-2030.

15

631. (subtilase) Streptomyces subtilisin-type inhibitors signature

Bacteria of the Streptomyces family produce a family of proteinase inhibitors[1] characterized by their strong activity toward subtilisin. They are collectively known as SSI's: Streptomyces Subtilisin Inhibitors. Some SSI's also inhibit trypsin or chymotrypsin. In their mature secreted form, SSI's are proteins of about 110 residues with two conserved disulfide bonds. +-----+ +-----+ |||

25 xxxxxxxxxxxxxxxCxxxxxxxxCxxxxxxxxCx#xxxxxxxxxxxxCxxxxxx *****'C': conserved cysteine involved in a disulfide bond.'#': active site residue.'*': position of the pattern.

Consensus pattern: C-x-P-x(2,3)-G-x-H-P-x(4)-A-C-[ATD]-x-L [The two C's are involved in a disulfide bond]

30 [1] Taguchi S., Kojima S., Terabe M., Miura K.-I., Momose H. Eur. J. Biochem. 220:911-918(1994).

632. Sugar transport proteins signatures

In mammalian cells the uptake of glucose is mediated by a family of closely related transport proteins which are called the glucose transporters [1,2,3]. At least seven of these transporters are currently known to exist (in Human they are encoded by the GLUT1 to GLUT7 genes). These integral membrane proteins are predicted to comprise twelve membrane spanning domains. The glucose transporters show sequence similarities [4,5] with a number of other sugar or metabolite transport proteins listed below (references are only provided for recently determined sequences). - *Escherichia coli* arabinose-proton symport (araE). - *Escherichia coli* galactose-proton symport (galP). - *Escherichia coli* and *Klebsiella pneumoniae* citrate-proton symport (also known as citrate utilization determinant) (gene cit). - *Escherichia coli* alpha-ketoglutarate permease (gene kgtP). - *Escherichia coli* proline/betaine transporter (gene proP) [6]. - *Escherichia coli* xylose-proton symport (xylE). - *Zymomonas mobilis* glucose facilitated diffusion protein (gene glf). - Yeast high and low affinity glucose transport proteins (genes SNF3, HXT1 to HXT14). - Yeast galactose transporter (gene GAL2). - Yeast maltose permeases (genes MAL3T and MAL6T). - Yeast myo-inositol transporters (genes ITR1 and ITR2). - Yeast carboxylic acid transporter protein homolog JEN1. - Yeast inorganic phosphate transporter (gene PHO84). - *Kluyveromyces lactis* lactose permease (gene LAC12). - *Neurospora crassa* quinate transporter (gene Qa-y), and *Emericella nidulans* quinate permease (gene qutD). - *Chlorella hexose* carrier (gene HUP1). - *Arabidopsis thaliana* glucose transporter (gene STP1). - Spinach sucrose transporter. - *Leishmania donovani* transporters D1 and D2. - *Leishmania enriettii* probable transport protein (LTP). - Yeast hypothetical proteins YBR241c, YCR98c and YFL040w. - *Caenorhabditis elegans* hypothetical protein ZK637.1. - *Escherichia coli* hypothetical proteins yabE, ydjE and yhjE. - *Haemophilus influenzae* hypothetical proteins HI0281 and HI0418. - *Bacillus subtilis* hypothetical proteins yxbC and yxdF. It has been suggested [4] that these transport proteins have evolved from the duplication of an ancestral protein with six transmembrane regions, this hypothesis is based on the conservation of two G-R-[KR] motifs. The first one is located between the second and third transmembrane domains and the second one between transmembrane domains 8 and 9. Two patterns have been developed to detect this family of proteins. The first pattern is based on the G-R-[KR] motif; but because this motif is too short to be specific to this family of proteins, a pattern from a larger region centered on the second copy of this motif was derived. The second pattern is based on a

number of conserved residues which are located at the end of the fourth transmembrane segment and in the short loop region between the fourth and fifth segments.

Consensus pattern: [LIVMSTAG]-[LIVMFSAG]-x(2)-[LIVMSA]-[DE]-x-[LIVMFYWA]-G- R-[RK]-x(4,6)-[GSTA]

5 Consensus pattern: [LIVMF]-x-G-[LIVMFA]-x(2)-G-x(8)-[LIFY]-x(2)-[EQ]-x(6)- [RK]
 [1] Silverman M. Annu. Rev. Biochem. 60:757-794(1991).[2] Gould G.W., Bell G.I. Trends Biochem. Sci. 15:18-23(1990).[3] Baldwin S.A. Biochim. Biophys. Acta 1154:17-49(1993).[4] Maiden M.C.J., Davis E.O., Baldwin S.A., Moore D.C.M., Henderson P.J.F. Nature 325:641-643(1987).[5] Henderson P.J.F. Curr. Opin. Struct. Biol. 1:590-601(1991).[6] Culham D.E., Lasby B., Marangoni A.G., Milner J.L., Steer B.A., van Nues R.W., Wood J.M. J. Mol. Biol. 229:268-276(1993).

633. Synaptobrevin signature

15 Synaptobrevin [1] is an intrinsic membrane protein of small synaptic vesicles whose function is not yet known, but which is highly conserved in mammals, electric ray (where its is known as VAMP-1), Drosophila and yeast [2]. In yeast there are two closely related forms of synaptobrevin (genes SNC1 and SNC2) while in mammals there is at least 4 (genes SYB1, SYB2, SYB3 and SYBL1). Structurally synaptobrevin consist of a N-terminal cytoplasmic domain of from 90 to 110 residues, followed by a transmembrane region, and then by a short (from 2 to 22 residues) C-terminal intravesicular domain. As a signature pattern for synaptobrevin, a highly conserved stretch of residues located in the central part of the sequence was selected.

20 Consensus pattern: N-[LIVM]-[DENS]-[KL]-V-x-[DEQ]-R-x(2)-[KR]-[LIVM]-[STDE]- x-[LIVM]-x-[DE]-[KR]-[TA]-[DE]

25 [1] Suedhof T.C., Baumert M., Perin M.S., Jahn R. Neuron 2:1475-1481(1989).[2] Gerst J.E., Rodgers L., Riggs M., Wigler M. Proc. Natl. Acad. Sci. U.S.A. 89:4338-4342(1992).

634. TBC domain. Identification of a TBC domain in GYP6_YEAST and GYP7_YEAST, which are GTPase activator proteins of yeast Ypt6 and Ypt7, imply that these domains are GTPase activator proteins of Rab-like small GTPases. Number of members: 55

[1] Medline: 96032578. Molecular cloning of a cDNA with a novel domain present in the tre-2 oncogene and the yeast cell cycle regulators BUB2 and cdc16. Richardson PM, Zon LI; Oncogene 1995;11:1139-1148.

[2]Medline: 97398935. A shared domain between a spindle assembly checkpoint protein and Ypt/Rab-specific GTPase-activators. Neuwald AF; Trends Biochem Sci 1997;22:243-244.

635. Transcription factor TFIID repeat signature (TBP)

Transcription factor TFIID (or TATA-binding protein, TBP) [1,2] is a general factor that plays a major role in the activation of eukaryotic genes transcribed by RNA polymerase II. TFIID binds specifically to the TATA box promoter element which lies close to the position of transcription initiation. There is a remarkable degree of sequence conservation of a C-terminal domain of about 180 residues in TFIID from various eukaryotic sources. This region is necessary and sufficient for TATA box binding. The most significant structural feature of this domain is the presence of two conserved repeats of a 77 amino-acid region. The intramolecular symmetry generates a saddle-shaped structure that sits astride the DNA [3]. Drosophila TRF (TBP-related factor) [4] is a sequence-specific transcription factor that also binds to the TATA box and is highly similar to TFIID. Archaeobacteria also possess a TBP homolog [5]. A signature pattern that spans the last 50 residues of the repeated region has been derived.-

Consensus pattern: Y-x-P-x(2)-[IF]-x(2)-[LIVM](2)-x-[KRH]-x(3)-P-[RKQ]-x(3)- L-[LIVM]-F-x-[STN]-G-[KR]-[LIVM]-x(3)-G-[TAGL]-[KR]-x(7)- [AGC]-x(7)-[LIVM]
 [1] Hoffmann A., Sinn E., Yamamoto T., Wang J., Roy A., Horikoshi M., Roeder R.G. Nature 346:387-390(1990).[2] Gash A., Hoffmann A., Horikoshi M., Roeder R.G., Chua N.-H. Nature 346:390-394(1990).[3] Nikolov D.B., Hu S.-H., Lin J., Gasch A., Hoffmann A., Horikoshi M., Chua N.-H., Roeder R.G., Burley S.K. Nature 360:40-46(1992).[4] Crowley T.E., Hoey T., Liu J.-K., Jan Y.N., Jan L.Y., Tjian R. Nature 361:557-561(1993).[5] Marsh T.L., Reich C.I., Whitelock R.B., Olsen G.J. Proc. Natl. Acad. Sci. U.S.A. 91:4180-4184(1994).

636. Translationally controlled tumor protein signatures (TCTP)

Mammalian translationally controlled tumor protein (TCTP) (or P23) is a protein which has been found to be preferentially synthesized in cells during the early growth phase of some types of tumor [1,2], but which is also expressed in normal cells. The physiological function of TCTP is still not known. It is a hydrophilic protein of 18 to 20 Kd. Close homologs have been found in plants [3], earthworm [4], *Caenorhabditis elegans* (F52H2.11), Hydra, budding yeast (YKL056c) [5] and fission yeast (SpAC1F12.02c). Two of the best conserved regions have been selected as signature patterns for TCTP.

Consensus pattern: [IFA]-[GA]-[GAS]-N-[PAK]-S-[GA]-E-[GDE]-[PAGE]-[DEQGA]

Consensus pattern: [FLVH]-[FY]-[IVCT]-G-E-x-[MA]-x(2,5)-[DEN]-[GAST]-x-[LV]-[AV]-x(3)-[FYW]

[1] Boehm H., Beendorf R., Gaestel M., Gross B., Nuernberg P., Kraft R., Otto A., Bielka H. *Biochem. Int.* 19:277-286(1989).[2] Makrides S., Chitpatima S.T., Bandyopadhyay R., Brawerman G. *Nucleic Acids Res.* 16:2350-2350(1988).[3] Pay A., Heberle-Bors E., Hirt H. *Plant Mol. Biol.* 19:501-503(1992).[4] Stuerzenbaum S.R., Kille P., Morgan A.J. *Biochim. Biophys. Acta* 1398:294-304(1998).[5] Rasmussen S.W. *Yeast* 10:S63-S68(1994).

637. TFIIS zinc ribbon domain signature

Transcription factor S-II (TFIIS) [1] is a eukaryotic protein necessary for efficient RNA polymerase II transcription elongation, past template-encoded pause sites. TFIIS shows DNA-binding activity only in the presence of RNA polymerase II. It is a protein of about 300 amino acids whose sequence is highly conserved in mammals, *Drosophila*, yeast (where it was first known as PPR2, a transcriptional regulator of URA4, and then as DST1, the DNA strand transfer protein alpha [2]) and in the archaebacteria *Sulfolobus acidocaldarius* [3]. This family also includes the eukaryotic and archebacterial RNA polymerase subunits of the 15 Kd / M family (see <PDOC00790>) as well as the following viral proteins: - Vaccinia virus RNA polymerase 30 Kd subunit (rpo30) [4]. - African swine fever virus protein I243L [5]. The best conserved region of all these proteins contains four cysteines that bind a zinc ion and fold in a conformation termed a 'zinc ribbon' [6]. Besides these cysteines, there are a number of other conserved residues which can be used to help define a specific pattern for this type of domain.

Consensus pattern: C-x(2)-C-x(9)-[LIVMQSAR]-[QH]-[STQL]-[RA]-[SACR]-x-[DE]-[DET]-[PGSEA]-x(6)-C-x(2,5)-C-x(3)-[FW] [The four C's are zinc ligands]

[1] Hirashima S., Hirai H., Nakanishi Y., Natori S. J. Biol. Chem. 263:3858-3863(1988).[2] Kipling D., Kearsey S.E. Nature 353:509-509(1991).[3] Langer D., Zillig W. Nucleic Acids Res. 21:2251-2251(1993).[4] Ahn B.-Y., Gershon P.D., Jones E.V., Moss B. Mol. Cell. Biol. 10:5433-5441(1990).[5] Rodriguez J.M., Salas M.L., Vinuela E. Virology 186:40-52(1992).[6] Qian X., Jeon C., Yoon H., Agarwal K., Weiss M.A. Nature 365:277-279(1993).

638. Tetrahydrofolate dehydrogenase/cyclohydrolase signatures (THF DHG CYH)

Enzymes that participate in the transfer of one-carbon units are involved in various biosynthetic pathways. In many of these processes the transfers of one-carbon units are mediated by the coenzyme tetrahydrofolate (THF). Various reactions generate one-carbon derivatives of THF which can be interconverted between different oxidation states by formyltetrahydrofolate synthetase(EC 6.3.4.3), methylenetetrahydrofolate dehydrogenase (EC 1.5.1.5 or EC 1.5.1.15) and methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9).The dehydrogenase and cyclohydrolase activities are expressed by a variety of multifunctional enzymes: - Eukaryotic C-1-tetrahydrofolate synthase (C1-THF synthase), which catalyzes all three reactions described above. Two forms of C1-THF synthases are known [1], one is located in the mitochondrial matrix, while the second one is cytoplasmic. In both forms the dehydrogenase/cyclohydrolase domain is located in the N-terminal section of the 900 amino acids protein and consists of about 300 amino acid residues. The C1-THF synthases are NADP- dependent. - Eukaryotic mitochondrial bifunctional dehydrogenase/cyclohydrolase [2]. This is an homodimeric NAD-dependent enzyme of about 300 amino acid residues. - Bacterial folD [3]. FolD is an homodimeric bifunctional NADP-dependent enzyme of about 290 amino acid residues. The sequence of the dehydrogenase/cyclohydrolase domain is highly conserved in all forms of the enzyme. Two conserved regions have been selected as signature patterns. The first one is located in the N-terminal part of these enzymes and contains three acidic residues. The second pattern is a highly conserved sequence of 9 amino acids which is located in the C-terminal section.

Consensus pattern: [EQ]-x-[EQK]-[LIVM](2)-x(2)-[LIVM]-x(2)-[LIVMY]-N-x-[DN]- x(5)-[LIVMF](3)-Q-L-P-[LV]

Consensus pattern: P-G-G-V-G-P-[MF]-T-[IV]

[1] Shannon K.W., Rabinowitz J.C. J. Biol. Chem. 263:7717-7725(1988).[2] Belanger C., Mackenzie R.E. J. Biol. Chem. 264:4837-4843(1989).[3] d'Ari L., Rabinowitz J.C. J. Biol. Chem. 266:23953-23958(1991).

5

639. Triosephosphate isomerase active site (TIM)

Triosephosphate isomerase (EC 5.3.1.1) (TIM) [1] is the glycolytic enzyme that catalyzes the reversible interconversion of glyceraldehyde 3-phosphate and dihydroxyacetone phosphate. TIM plays an important role in several metabolic pathways and is essential for efficient energy production. It is a dimer of identical subunits, each of which is made up of about 250 amino-acid residues. A glutamic acid residue is involved in the catalytic mechanism [2]. The sequence around the active site residue is perfectly conserved in all known TIM's and can be used as a signature pattern for this type of enzyme.

10

Consensus pattern: [AV]-Y-E-P-[LIVM]-W-[SA]-I-G-T-[GK] [E is the active site residue]
[1] Lolis E., Alber T., Davenport R.C., Rose D., Hartman F.C., Petsko G.A. Biochemistry 29:6609-6618(1990).[2] Knowles J.R. Nature 350:121-124(1991).

5

640. Thymidine kinase cellular-type signature (TK)

Thymidine kinase (TK) (EC 2.7.1.21) is an ubiquitous enzyme that catalyzes the ATP-dependent phosphorylation of thymidine. A comparison of TK sequences has shown [1,2,3] that there are two different families of TK. One family groups together TK from herpes viruses as well as cellular thymidylate kinases, while the second family currently consists of TK from the following sources: - Vertebrates. - Bacterial. - Bacteriophage T4. - Pox viruses. - African swine fever virus (ASF). - Fish lymphocystis disease virus (FLDV). A conserved region which is located in the C-terminal section of these enzymes has been selected as a signature pattern for this family of TKA.

25

Consensus pattern: [GA]-x(1,2)-[DE]-x-Y-x-[STAP]-x-C-[NKR]-x-[CH]-[LIVMFYWH]
[1] Boyle D.B., Coupar B.E.H., Gibbs A.J., Seigman L.J., Both G.W. Virology 156:355-365(1987).[2] Blasco R., Lopez-Otin C., Munoz M., Bockamp E.-O., Simon-Mateo C., Vinuela E. Virology 178:301-304(1990).[3] Robertson G.R., Whalley J.M. Nucleic Acids Res. 16:11303-11317(1988).

30

641. Thymidine kinase from herpesvirus (TK herpes)

[1]

Medline: 96003730

- 5 Crystal structures of the thymidine kinase from herpes simplex virus type-1 in complex with deoxythymidine and ganciclovir.

Brown DG, Visse R, Sandhu G, Davies A, Rizkallah PJ, Melitz C, Summers WC, Sanderson MR;

10 Nat Struct Biol 1995;2:876-881.

Number of members: 65

642. Nuclear transition protein 2 signatures (TP2)

15 In mammals, the second stage of spermatogenesis is characterized by the conversion of nucleosomal chromatin to the compact, non-nucleosomal and transcriptionally inactive form found in the sperm nucleus. This condensation is associated with a double-protein transition. The first transition corresponds to the replacement of histones by several spermatid-specific proteins, also called transition proteins, which are themselves replaced by protamines during the second transition. Nuclear transition protein 2 (TP2) is one of those spermatid-specific proteins. TP2 is a basic, zinc-binding protein [1] of 116 to 137 amino-acid residues.

20 Structurally, TP2 consists of three distinct parts: a conserved serine-rich N-terminal domain of about 25 residues, a variable central domain of 20 to 50 residues which contains cysteine residues, and a conserved C-terminal domain of about 70 residues rich in lysines and
25 arginines. Two signature patterns for TP2 have been developed: one located in the N-terminal domain, the other in the C-terminal.

Consensus pattern: H-x(3)-H-S-[NS]-S-x-P-Q-S

Consensus pattern: K-x-R-K-x(2)-E-G-K-x(2)-K-[KR]-K

[1] Baskaran R., Rao M.R.S. Biochem. Biophys. Res. Commun. 179:1491-1499(1991).

643. Thiamine pyrophosphate enzymes signature (TTP enzymes)

A number of enzymes require thiamine pyrophosphate (TPP) (vitamin B1) as a cofactor. It has been shown [1] that some of these enzymes are structurally related. These related TPP enzymes are: - Pyruvate oxidase (POX) (EC 1.2.3.3) Reaction catalyzed: pyruvate + orthophosphate + O(2) + H(2)O = acetyl phosphate + CO(2) + H(2)O(2). - Pyruvate decarboxylase (PDC) (EC 4.1.1.1) Reaction catalyzed: pyruvate = acetaldehyde + CO(2). - Indolepyruvate decarboxylase (EC 4.1.1.74) [2] Reaction catalyzed: indole-3-pyruvate = indole-3-acetaldehyde + CO(2). - Acetolactate synthase (ALS) (EC 4.1.3.18) Reaction catalyzed: 2 pyruvate = acetolactate + CO(2). - Benzoylformate decarboxylase (BFD) (EC 4.1.1.7) [3] Reaction catalyzed: benzoylformate = benzaldehyde + CO(2). A conserved region which is located in their C-terminal section has been selected as a signature pattern for these enzymes.

Consensus pattern: [LIVMF]-[GSA]-x(5)-P-x(4)-[LIVMFYW]-x-[LIVMF]-x-G-D-[GSA]-[GSAC]

[1] Green J.B.A. FEBS Lett. 246:1-5(1989).[2] Koga J., Adachi T., Hidaka H. Mol. Gen. Genet. 226:10-16(1991).[3] Tsou A.Y., Ransom S.C., Gerlt J.A., Buechter D.D., Babbitt P.C., Kenyon G.L. Biochemistry 29:9856-9862(1990).

644. TPR Domain

[1]

Medline: 95397415

Tetratrico peptide repeat interactions: to TPR or not to TPR?

Lamb JR, Tugendreich S, Hieter P;

Trends Biochem Sci 1995;20:257-259.

[2]Medline: 98151343

The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for TPR-mediated protein-protein interactions.

Das AK, Cohen PW, Barford D;

EMBO J 1998;17:1192-1199.

Number of members: 621

645. Uroporphyrin-III C-methyltransferase signatures (TP methylase)

Uroporphyrin-III C-methyltransferase (EC 2.1.1.107) (SUMT) [1,2] catalyzes the transfer of two methyl groups from S-adenosyl-L-methionine to the C-2 and C-7 atoms of uroporphyrinogen III to yield precorrin-2 via the intermediate formation of precorrin-1.

SUMT is the first enzyme specific to the cobalamin pathway and precorrin-2 is a common intermediate in the biosynthesis of corrinoids such as vitamin B12, siroheme and coenzyme F430. The sequences of SUMT from a variety of eubacterial and archaeobacterial species are currently available. In species such as *Bacillus megaterium* (gene *cobA*), *Pseudomonas denitrificans* (*cobA*) or *Methanobacterium ivanovii* (gene *corA*) SUMT is a protein of about 25 to 30 Kd. In *Escherichia coli* and related bacteria, the *cysG* protein, which is involved in the biosynthesis of siroheme, is a multifunctional protein composed of a N-terminal domain, probably involved in transforming precorrin-2 into siroheme, and a C-terminal domain which has SUMT activity. The sequence of SUMT is related to that of a number of *P. denitrificans* and *Salmonella typhimurium* enzymes involved in the biosynthesis of cobalamin which also seem to be SAM-dependent methyltransferases [3,4]. The similarity is especially strong with two of these enzymes: *cobI/cbiL* which encodes S-adenosyl-L-methionine--precorrin-2 methyltransferase and *cobM/cbiF* whose exact function is not known. Two signature patterns have been developed for these enzymes. The first corresponds to a well conserved region in the N-terminal extremity (called region 1 in [1,3]) and the second to a less conserved region located in the central part of these proteins (this pattern spans what are called regions 2 and 3 in [1,3]).

Consensus pattern: [LIVM]-[GS]-[STAL]-G-P-G-x(3)-[LIVMFY]-[LIVM]-T-[LIVM]-[KRHOG]-[AG]

Consensus pattern: V-x(2)-[LI]-x(2)-G-D-x(3)-[FYW]-[GS]-x(8)-[LIVF]-x(5,6)-

[LIVMFYWPAC]-x-[LIVMY]-x-P-G

[1] Blanche F., Robin C., Couder M., Faucher D., Cauchois L., Cameron B., Crouzet J. J. Bacteriol. 173:4637-4645(1991).[2] Robin C., Blanche F., Cauchois L., Cameron B., Couder M., Crouzet J. J. Bacteriol. 173:4893-4896(1991).[3] Crouzet J., Cameron B., Cauchois L., Rigault S., Rouyez M.-C., Blanche F., Thibaut D., Debussche L. J. Bacteriol. 172:5980-5990(1990).[4] Roth J.R., Lawrence J.G., Rubenfield M., Kieffer-Higgins S., Church G.M. J. Bacteriol. 175:3303-3316(1993).[5] Mattheakis L.C., Shen W.H., Collier R.J. Mol. Cell. Biol. 12:4026-4037(1992).

646. Tudor domain

Domain of unknown function present in several RNA-binding proteins. copies in the Drosophila Tudor protein. Slight ambiguities in the alignment. Number of members: 18

[1]Medline: 97200561 Tudor domains in proteins that interact with RNA. Ponting CP; Trends Biochem Sci 1997;22:51-52. [2]Medline: 97157029 The human EBNA-2 coactivator p100: multidomain organization and relationship to the staphylococcal nuclease fold and to the tudor protein involved in Drosophila melanogaster development. Callebaut I, Mornon JP; Biochem J 1997;321:125-132.

647. Terpene synthase family

It has been suggested that this gene family be designated tps (for terpene synthase) [1]. It has been split into six subgroups on the basis of phylogeny, called tpsa-tpsf.

tpsa includes vetispiradiene synthase Swiss:Q39979, 5-epi-aristolochene synthase, Swiss:Q40577 and (+)-delta-cadinene synthase Swiss:P93665.

tpsb includes (-)-limonene synthase, Swiss:Q40322.

tpsc includes kaurene synthase A, Swiss:O04408.

tpsd includes taxadiene synthase, Swiss:Q41594, pinene synthase, Swiss:O24475 and myrcene synthase, Swiss:O24474.

tpse includes kaurene synthase B.

tpsf includes linalool synthase.

Number of members: 51

[1]

Medline: 97413772

Monoterpene synthases from grand fir (*Abies grandis*). cDNA isolation, characterization, and functional expression of myrcene synthase, (-)-(4S)-limonene synthase, and (-)-(1S,5S)-pinene synthase.

Bohlmann J, Steele CL, Croteau R;

J Biol Chem 1997;272:21784-21792.

648. ThiF family

This family contains a repeated domain in ubiquitin
activating enzyme E1 and members of the bacterial
ThiF/MoeB/HesA family. Number of members: 87

649. Thioester dehydrase

Members of this family are involved in fatty acid biosynthesis.
Number of members: 19

[1]

Medline: 96398612

Structure of a dehydratase-isomerase from the bacterial
pathway for biosynthesis of unsaturated fatty acids: two
catalytic activities in one active site.

Leesong M, Henderson BS, Gillig JR, Schwab JM, Smith JL;
Structure 1996;4:253-264.

Database Reference: SCOP; 1mka; fa; [SCOP-USA][CATH-PDBSUM]

Database reference: PFAMB; PB058036;

650. Tub family signatures

The mouse tubby mutation is the cause of maturity-onset obesity, insulin resistance and
sensory deficits. This mutation maps to a gene, tub [1,2], which codes for a protein that
belongs to a family which currently consists of the following members: - Mammalian tub, an
hydrophilic protein of about 500 residues, which could be involved in the hypothalamic
regulation of body weight. - Human protein TULP1 [3] which may be involved in retinis
pigmentosa 14, a retinal degeneration disease. - Mouse protein p4-6 whose function is not
known. - Caenorhabditis elegans hypothetical protein F10B5.4. - Several fragmentary
sequences from plants, Drosophila and human ESTs. While the N-terminal part of these
protein is not conserved in length nor in the sequence, the C-terminal 250 residues are highly
conserved. Therefore, two regions were selected in the C-terminal part as signature patterns.

The second region is located at the C-terminal extremity and contains a penultimate cysteine residue that could be critical to the normal functioning of these proteins.

Consensus pattern: F-[KHQ]-G-R-V-[ST]-x-A-S-V-K-N-F-Q

Consensus pattern: A-F-[AG]-I-[SAC]-[LIVM]-[ST]-S-F-x-[GST]-K-x-A-C-E

- 5 [1] Kleyn P.W., Fan W., Kovats S.G., Lee J.L., Pulido J.C., Wu Y., Berkemeier L.R., Misumi D.J., Holmgren L., Charlat O., Woolf E.A., Tayber O., Brody T., Shu P., Hawkins F., Kennedy B., Baldini L., Ebeling C., Alperin G.D., Deeds J., Lakey N.D., Culpepper J., Chen H., Gluecksmann-Kuis M.A., Carlson G.A., Duyk G.M., Moore K.J. *Cell* 85:281-290(1996).[2] Noben-Trauth K., Naggert J.K., North M.A., Nishina P.M. *Nature* 380:534-538(1996).[3] North M.A., Naggert J.K., Yan Y., Noben-Trauth K., Nishina P.M. *Proc. Natl. Acad. Sci. U.S.A.* 94:3128-3133(1997).

651. Eukaryotic DNA topoisomerase I active site

15 DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type I topoisomerases act by catalyzing the transient breakage of DNA, one strand at a time, and the subsequent rejoining of the strands. When a eukaryotic type I topoisomerase breaks a DNA backbone bond, it simultaneously forms a protein-DNA link where the hydroxyl group of a tyrosine residue is joined to a 3'-phosphate on DNA, at one end of the enzyme-severed DNA strand. In 20 eukaryotes and pox virus topoisomerases I, there are a number of conserved residues in the region around the active site tyrosine.

Consensus pattern: [DEN]-x(6)-[GS]-[IT]-S-K-x(2)-Y-[LIVM]-x(3)-[LIVM] [Y is the active site tyrosine]

- 25 [1] Sternglanz R. *Curr. Opin. Cell Biol.* 1:533-535(1990).[2] Sharma A., Mondragon A. *Curr. Opin. Struct. Biol.* 5:39-47(1995).[3] Lynn R.M., Bjornsti M.-A., Caron P.R., Wang J.C. *Proc. Natl. Acad. Sci. U.S.A.* 86:3559-3563(1989).[4] Roca J. *Trends Biochem. Sci.* 20:156-160(1995).[E1]

30

652. Transaldolase signatures

Transaldolase (EC 2.2.1.2) catalyzes the reversible transfer of a three-carbon ketol unit from sedoheptulose 7-phosphate to glyceraldehyde 3-phosphate to form erythrose 4-phosphate and

fructose 6-phosphate. This enzyme, together with transketolase, provides a link between the glycolytic and pentose-phosphate pathways. Transaldolase is an enzyme of about 34 Kd whose sequence has been well conserved throughout evolution. A lysine has been implicated [1] in the catalytic mechanism of the enzyme; it acts as a nucleophilic group that attacks the carbonyl group of fructose-6-phosphate. Transaldolase is evolutionary related [2] to a bacterial protein of about 20Kd (known as talC in *Escherichia coli*), whose exact function is not yet known. Two signature patterns have been developed for these proteins. The first, located in the N-terminal section, contains a perfectly conserved pentapeptide; these cond, includes the active site lysine.

Consensus pattern: [DG]-[IVSA]-T-[ST]-N-P-[STA]-[LIVMF](2)

Consensus pattern: [LIVM]-x-[LIVM]-K-[LIVM]-[PAS]-x-[ST]-x-[DENQPAS]-G-[LIVM]-x-[AGV]-x-[QEKRTS]-x-[LIVM] [K is the active site residue]

[1] Miosga T., Schaaff-Gerstenschlaeger I., Franken E., Zimmermann F.K. Yeast 9:1241-1249(1993).[2] Reizer J., Reizer A., Saier M.H. Jr. Microbiology 141:961-971(1995).

653. (Transpeptidase) Penicillin binding protein transpeptidase domain

The active site serine (residue 337 in [Swiss:P14677](#)) is conserved in all members of this family.

[1] Pares S, Mouz N, Petillot Y, Hakenbeck R, Dideberg O Nat Struct Biol 1996;3:284-289.

654. Trehalase signatures

Trehalase (EC [3.2.1.28](#)) is the enzyme responsible for the degradation of the disaccharide alpha, alpha-trehalose yielding two glucose subunits [1]. It is an enzyme found in a wide variety of organisms and whose sequence has been highly conserved throughout evolution. Two of the most highly conserved regions have been selected as signature patterns. The first pattern is located in the central section, the second one is in the C-terminal region.

Consensus pattern: P-G-G-R-F-x-E-x-Y-x-W-D-x-Y

Consensus pattern: Q-W-D-x-P-x-[GA]-W-[PAS]-P

[1] Kopp M., Mueller H., Holzer H. J. Biol. Chem. 268:4766-4774(1993).[2] Henrissat B., Bairoch A. Biochem. J. 293:781-788(1993).[E1]

5 655. Trehalose-6-phosphate synthase domain

OtsA (Trehalose-6-phosphate synthase) is homologous to regions in the subunits of yeast trehalose-6-phosphate synthase/phosphate complex, [1].

[1] Kaasen I, McDougall J, Strom AR; Gene 1994;145:9-15.

10

656. Tropomyosins signature

Tropomyosins [1,2] are family of closely related proteins present in muscle and non-muscle cells. In striated muscle, tropomyosin mediate the interactions between the troponin complex and actin so as to regulate muscle contraction. The role of tropomyosin in smooth muscle and non-muscle tissues is not clear. Tropomyosin is an alpha-helical protein that forms a coiled-coil dimer. Muscle isoforms of tropomyosin are characterized by having 284 amino acid residues and a highly conserved N-terminal region, whereas non-muscle forms are generally smaller and are heterogeneous in their N-terminal region. The signature pattern for tropomyosins is based on a very conserved region in the C-terminal section of tropomyosins and which is present in both muscle and non-muscle forms.

Consensus pattern: L-K-E-A-E-x-R-A-E

[1] Smilie L.B. Trends Biochem. Sci. 4:151-155(1979).[2] McLeod A.R. BioEssays 6:208-212(1986).

25

657. Troponin

Troponin (Tn) contains three subunits, Ca²⁺ binding (TnC), inhibitory (TnI), and tropomyosin binding (TnT). this Pfam contains members of the TnT subunit.

30

Troponin is a complex of three proteins, Ca²⁺ binding (TnC), inhibitory (TnI), and tropomyosin binding (TnT).

The troponin complex regulates Ca⁺⁺ induced muscle contraction.

This family includes troponin T and troponin I. Troponin I

binds to actin and troponin T binds to tropomyosin.

Number of members: 81 [1]

Medline: 87144593

Structure of co-crystals of tropomyosin and troponin.

- 5 White SP, Cohen C, Phillips GN Jr;
Nature 1987;325:826-828. [2]

Medline: 95155315

A direct regulatory role for troponin T and a dual role for
troponin C in the Ca²⁺ regulation of muscle contraction.

- 10 Potter JD, Sheng Z, Pan BS, Zhao J;
J Biol Chem 1995;270:2557-2562.
[3]Medline: 95324796

The troponin complex and regulation of muscle contraction.

- Farah CS, Reinach FC;
FASEB J 1995;9:755-767.

658. (Tryp mucin) Mucin-like glycoprotein

20 This family of trypanosomal proteins resemble vertebrate mucins. The protein consists of
three regions. The N and C terminii are conserved between all members of the family,
whereas the central region is not well conserved and contains a large number of threonine
residues which can be glycosylated [1].

Indirect evidence suggested that these genes might encode the core protein of parasite
25 mucins, glycoproteins that were proposed to be involved in the interaction with, and invasion
of, mammalian host cells.

[1] Di Noia JM, Sanchez DO, Frasch AC; J Biol Chem 1995;270:24146-24149.

[2] Di Noia JM, D'Orso I, Aslund L, Sanchez DO, Frasch AC; J Biol Chem 1998;273:10843-
30 10850.

659. Aminoacyl-transfer RNA synthetases class-I signature (tRNA synt 1)

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. A few years ago it was found [2] that several aminoacyl-tRNA synthetases share a region of similarity in their N-terminal section, in particular the consensus tetrapeptide His-Ile-Gly-His ('HIGH') is very well conserved. The 'HIGH' region has been shown [3] to be part of the adenylate binding site. The 'HIGH' signature has been found in the aminoacyl-tRNA synthetases specific for arginine, cysteine, glutamic acid, glutamine, isoleucine, leucine, methionine, tyrosine, tryptophan, and valine. These aminoacyl-tRNA synthetases are referred to as class-I synthetases [4,5,6] and seem to share the same tertiary structure based on a Rossmann fold. Consensus pattern: P-x(0,2)-[GSTAN]-[DENQGAPK]-x-[LIVMFP]-[HT]-[LIVMYAC]-G-[HNTG]-[LIVMFYSTAGPC]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Webster T., Tsai H., Kula M., Mackie G.A., Schimmel P. Science 226:1315-1317(1984).[3] Brick P., Bhat T.N., Blow D.M. J. Mol. Biol. 208:83-98(1988).[4] Delarue M., Moras D. BioEssays 15:675-687(1993).[5] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[6] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

660. Aminoacyl-transfer RNA synthetases class-I signature (tRNA synt 1b)

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. A few years ago it was found [2] that several aminoacyl-tRNA synthetases share a region of similarity in their N-terminal section, in particular the consensus tetrapeptide His-Ile-Gly-His ('HIGH') is very well

conserved. The 'HIGH' region has been shown [3] to be part of the adenylate binding site.

The 'HIGH' signature has been found in the aminoacyl-tRNA synthetases specific

for arginine, cysteine, glutamic acid, glutamine, isoleucine, leucine, methionine, tyrosine,

tryptophan, and valine. These aminoacyl-tRNA synthetases are referred to as class-I

synthetases [4,5,6] and seem to share the same tertiary structure based on a Rossmann fold.

Consensus pattern: P-x(0,2)-[GSTAN]-[DENQGAPK]-x-[LIVMFP]-[HT]-[LIVMYAC]-G-[HNTG]-[LIVMFYSTAGPC]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Webster T., Tsai H., Kula M.,

Mackie G.A., Schimmel P. Science 226:1315-1317(1984).[3] Brick P., Bhat T.N., Blow

D.M. J. Mol. Biol. 208:83-98(1988).[4] Delarue M., Moras D. BioEssays 15:675-

687(1993).[5] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[6] Nagel G.M., Doolittle

R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

661. (tRNA-synt 1C) tRNA synthetases class I (E and Q)

Other tRNA synthetase sub-families are too dissimilar to be included.

This family includes only glutamyl and glutaminyl tRNA synthetases.

In some organisms, a single glutamyl-tRNA synthetase aminoacylates both tRNA(Glu) and tRNA(Gln).

[1] Rath VL, Silvian LF, Beijer B, Sproat BS, Steitz TA; Structure 1998;6:439-449.

662. (tRNA-synt 1d) tRNA synthetases class I (R)

Other tRNA synthetase sub-families are too dissimilar to be included.

This family includes only arginyl tRNA synthetase.

663. Aminoacyl-transfer RNA synthetases class-II signatures (tRNA synt 2)

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In

prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely

5 diverse in terms of subunit size and of quaternary structure. The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is

10 different to the Rossmann fold observed for the class I synthetases [7]. Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns have been derived from two of these regions.

Consensus pattern: [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]

Consensus pattern: [GSTALVF]-{DENQHRKP}-[GSTA]-[LIVMF]-[DE]-R-[LIVMF]-x-[LIVMSTAG]-[LIVMFY]

15 [1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Delarue M., Moras D. BioEssays 15:675-687(1993).[3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[4] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991). [5] Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).[6] Cusack S. Biochimie 75:1077-1081(1993).[7] Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-255(1990).[8] Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).

664. Aminoacyl-transfer RNA synthetases class-I signature (tRNA synt 1e)

25 Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a

30 mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. A few years ago it was found [2] that several aminoacyl-tRNA synthetases share a region of similarity in their N-terminal section, in particular the consensus tetrapeptide His-Ile-Gly-His ('HIGH') is very well

conserved. The 'HIGH' region has been shown [3] to be part of the adenylate binding site.

The 'HIGH' signature has been found in the aminoacyl-tRNA synthetases specific for arginine, cysteine, glutamic acid, glutamine, isoleucine, leucine, methionine, tyrosine, tryptophan, and valine. These aminoacyl-tRNA synthetases are referred to as class-I

synthetases [4,5,6] and seem to share the same tertiary structure based on a Rossmann fold.

Consensus pattern: P-x(0,2)-[GSTAN]-[DENQGAPK]-x-[LIVMFP]-[HT]-[LIVMYAC]-G-[HNTG]-[LIVMFYSTAGPC]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Webster T., Tsai H., Kula M., Mackie G.A., Schimmel P. Science 226:1315-1317(1984).[3] Brick P., Bhat T.N., Blow D.M. J. Mol. Biol. 208:83-98(1988).[4] Delarue M., Moras D. BioEssays 15:675-687(1993).[5] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[6] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

665. Aminoacyl-transfer RNA synthetases class-II signatures (tRNA synt 2b)

Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is different to the Rossmann fold observed for the class I synthetases [7]. Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns have been derived from two of these regions.

Consensus pattern: [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]

Consensus pattern: [GSTALVF]-{DENQHRKP}-[GSTA]-[LIVMF]-[DE]-R-[LIVMF]-x-[LIVMSTAG]-[LIVMFY]

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Delarue M., Moras D. BioEssays 15:675-687(1993).[3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[4] Nagel

G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991). [5] Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).[6] Cusack S. Biochimie 75:1077-1081(1993).[7] Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-255(1990).[8] Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).

666. Thaumatin family signature

Thaumatococcus daniellii, an African brush. The protein is made of about 200 residues and contains 8 disulfide bonds. A number of proteins have been found to be related to thaumatins. These protein are listed below (references are only provided for recently determined sequences). - A maize alpha-amylase/trypsin inhibitor. - Two tobacco pathogenesis-related proteins: PR-R major and minor forms, which are induced after infection with viruses. - Salt-induced protein NP24 from tomato. - Osmotin, a salt-induced protein from tobacco. - Osmotin-like proteins OSML13, OSML15 and OSML81 from potato [2]. - P21, a leaf protein from soybean. - PWIR2, a leaf protein from wheat. - Zeamatin, a maize antifungal protein [3]. The exact biological function of all these proteins is not yet known. A conserved region that includes three cysteine residues known (in thaumatococcus) to be involved in disulfide bonds has been selected as a signature pattern.

```
+-----+ | +-----+ | | ***** |
||
xxCxxxxxxxxxxxxxxxxCxxCxxCxxxxxxxxxxxxxxxxCxxCxCxxxCxCxxCCxCxxxCxxxxxC
xxxCx ||||| || | +-+ +-+ | +---+ +---+-+ | +-----+'C': conserved cysteine
involved in a disulfide bond.'*': position of the pattern.
```

Consensus pattern: G-x-[GF]-x-C-x-T-[GA]-D-C-x(1,2)-G-x(2,3)-C

[1] Edens L., Heslinga L., Klok R., Ledebroer A.M., Maat J., Toonen M.Y., Visser C., Verrips C.T. Gene 18:1-12(1982).[2] Zhu B., Chen T.H.H., Li P.H. Plant Physiol. 108:929-937(1995).[3] Malehorn D.E., Borgmeyer J.R., Smith C.E., Shah D.M.; Plant Physiol. 106:1471-1481(1994).

667. Thiolases signatures

Two different types of thiolase [1,2,3] are found both in eukaryotes and in prokaryotes: acetoacetyl-CoA thiolase (EC 2.3.1.9) and 3-ketoacyl-CoA thiolase (EC 2.3.1.16). 3-ketoacyl-CoA thiolase (also called thiolase I) has a broad chain-length specificity for its substrates and is involved in degradative pathways such as fatty acid beta-oxidation. Acetoacetyl-CoA

5 thiolase (also called thiolase II) is specific for the thiolysis of acetoacetyl-CoA and involved in biosynthetic pathways such as poly beta-hydroxybutyrate synthesis or steroid biogenesis. In eukaryotes, there are two forms of 3-ketoacyl-CoA thiolase: one located in the mitochondrion and the other in peroxisomes. There are two conserved cysteine residues important for thiolase activity. The first located in the N-terminal section of the enzymes is involved in the

10 formation of an acyl-enzyme intermediate; the second located at the C-terminal extremity is the active site base involved in deprotonation in the condensation reaction. Mammalian nonspecific lipid-transfer protein (nsL-TP) (also known as sterol carrier protein 2) is a protein which seems to exist in two different forms: a 14 Kd protein (SCP-2) and a larger 58 Kd protein (SCP-x). The former is found in the cytoplasm or the mitochondria and is involved in

15 lipid transport; the latter is found in peroxisomes. The C-terminal part of SCP-x is identical to SCP-2 while the N-terminal portion is evolutionary related to thiolases[4]. Three signature patterns have been developed for this family of proteins, two of which are based on the regions around the biologically important cysteines. The third is based on a highly conserved region in the C-terminal part of these proteins.

20 Consensus pattern: [LIVM]-[NST]-x(2)-C-[SAGLI]-[ST]-[SAG]-[LIVMFYNS]-x- [STAG]-[LIVM]-x(6)-[LIVM] [C is involved in formation of acyl-enzyme intermediate]

Consensus pattern: N-x(2)-G-G-x-[LIVM]-[SA]-x-G-H-P-x-[GA]-x-[ST]-G

Consensus pattern: [AG]-[LIVMA]-[STAGCLIVM]-[STAG]-[LIVMA]-C-x-[AG]-x-[AG]-x- [AG]-x-[SAG] [C is the active site residue]

25 [1] Peoples O.P., Sinskey A.J. J. Biol. Chem. 264:15293-15297(1989).[2] Yang S.-Y., Yang X.-Y.H., Healy-Louie G., Schulz H., Elzinga M. J. Biol. Chem. 265:10424-10429(1990).[3] Igual J.C., Gonzalez-Bosch C., Dopazo J., Perez-Ortin J.E. J. Mol. Evol. 35:147-155(1992).[4] Baker M.E., Billheimer J.T., Strauss J.F. III DNA Cell Biol. 10:695-698(1991).

668. Thioredoxin family active site

Thioredoxins [1 to 4] are small proteins of approximately one hundred amino-acid residues which participate in various redox reactions via the reversible oxidation of an active center

disulfide bond. They exist in either a reduced form or an oxidized form where the two cysteine residues are linked in an intramolecular disulfide bond. Thioredoxin is present in prokaryotes and eukaryotes and the sequence around the redox-active disulfide bond is well conserved. Bacteriophage T4 also encodes for a thioredoxin but its primary structure is not homologous to bacterial, plant and vertebrate thioredoxins. A number of eukaryotic proteins contain domains evolutionary related to thioredoxin, all of them seem to be protein disulphide isomerases (PDI). PDI (EC 5.3.4.1) [5,6,7] is an endoplasmic reticulum enzyme that catalyzes the rearrangement of disulfide bonds in various proteins. The various forms of PDI which are currently known are: - PDI major isozyme; a multifunctional protein that also function as the beta subunit of prolyl 4-hydroxylase (EC 1.14.11.2), as a component of oligosaccharyl transferase (EC 2.4.1.119), as thyroxine deiodinase (EC 3.8. 1.4), as glutathione-insulin transhydrogenase (EC 1.8.4.2) and as a thyroid hormone-binding protein ! - ERp60 (ER-60; 58 Kd microsomal protein). ERp60 was originally thought to be a phosphoinositide-specific phospholipase C isozyme and later to be a protease. - ERp72. - P5. All PDI contains two or three (ERp72) copies of the thioredoxin domain. Bacterial proteins that act as thiol:disulfide interchange proteins that allow disulfide bond formation in some periplasmic proteins also contain a thioredoxin domain. These proteins are: - Escherichia coli dsbA (or prfA) and its orthologs in Vibrio cholerae (tcpG) and Haemophilus influenzae (por). - Escherichia coli dsbC (or xpRA) and its orthologs in Erwinia chrysanthemi and Haemophilus influenzae. - Escherichia coli dsbD (or dipZ) and its Haemophilus influenzae ortholog. - Escherichia coli dsbE (or ccmG) and orthologs in Haemophilus influenzae, Rhodobacter capsulatus (helX), Rhizobiaceae (cycY and tlpA). Consensus pattern: [LIVMF]-[LIVMSTA]-x-[LIVMFYC]-[FYWSTHE]-x(2)-[FYWGNTN]-C-[GATPLVE]-[PHYWSTA]-C-x(6)-[LIVMFYWT] [The two C's form the redox-active bond]

[1] Holmgren A. Annu. Rev. Biochem. 54:237-271(1985). [2] Gleason F.K., Holmgren A. FEMS Microbiol. Rev. 54:271-297(1988). [3] Holmgren A. J. Biol. Chem. 264:13963-13966(1989). [4] Eklund H., Gleason F.K., Holmgren A. Proteins 11:13-28(1991). [5] Freedman R.B., Hawkins H.C., Murrant S.J., Reid L. Biochem. Soc. Trans. 16:96-99(1988). [6] Kivirikko K.I., Myllylä R., Pihlajaniemi T. FASEB J. 3:1609-1617(1989). [7] Freedman R.B., Hirst T.R., Tuite M.F. Trends Biochem. Sci. 19:331-336(1994).

669. (Transcript fac2) Transcription factor TFIIB repeat signature

In eukaryotes the initiation of transcription of protein encoding genes by polymerase II is modulated by general and specific transcription factors. The general transcription factors operate through common promoters elements (such as the TATA box). At least seven different proteins associates to form the general transcription factors: TFIIA, -IIB, -IID, -IIE, -IIF, -IIG, and -IIH[1]. Transcription factor IIB (TFIIB) plays a central role in the transcription of class II genes, it associates with a complex of TFIID-IIA bound to DNA (DA complex) to form a ternary complex TFIID-IIA-IBB (DAB complex) which is then recognized by RNA polymerase II [2,3]. TFIIB is a protein of about 315 to 340 amino acid residues which contains, in its C-terminal part an imperfect repeat of a domain of about 75 residues. This repeat could contribute an element of symmetry to the folded protein. The following proteins have been shown to be evolutionary related to TFIIB: - An archaebacterial TFIIB homolog. In *Pyrococcus woesei* a previously undetected open reading frame has been shown [4] to be highly related to TFIIB. - Fungal transcription factor IIIB 70 Kd subunit (gene PCF4/TDS4/BRF1) [5]. This protein is a general activator of RNA polymerase III transcription and plays a role analogous to that of TFIIB in pol III transcription. The central section of the repeated domain, which is the most conserved part of that domain has been selected as a signature pattern.

Consensus pattern: G-[KR]-x(3)-[STAGN]-x-[LIVMYA]-[GSTA](2)-[CSAV]-[LIVM]-[LIVMFY]-[LIVMA]-[GSA]-[STAC]

[1] Weinmann R. *Gene Expr.* 2:81-91(1992).[2] Hawley D. *Trends Biochem. Sci.* 16:317-318(1991).[3] Ha I., Lane W.S., Reinberg D. *Nature* 352:689-695(1991).[4] Ouzounis C., Sander C. *Cell* 71:189-190(1992).[5] Khoo B., Brophy B., Jackson S.P. *Genes Dev.* 8:2879-2890(1994).

670. (transcript fact) MADS-box domain signature and profile

A number of transcription factors contain a conserved domain of 56 amino-acid residues, sometimes known as the MADS-box domain [E1]. They are listed below: - Serum response factor (SRF) [1], a mammalian transcription factor that binds to the Serum Response Element (SRE). This is a short sequence of dyad symmetry located 300 bp to the 5' end of the transcription initiation site of genes such as c-fos. - Mammalian myocyte-specific enhancer factors 2A to 2D (MEF2A to MEF2D). These proteins are transcription factor which binds

specifically to the MEF2 element present in the regulatory regions of many muscle-specific genes. - *Drosophila* myocyte-specific enhancer factor 2 (MEF2). - Yeast GRM/PRTF protein (gene MCM1) [2], a transcriptional regulator of mating-type-specific genes. - Yeast arginine metabolism regulation protein I (gene ARGR1 or ARG80). - Yeast transcription factor

5 RLM1. - Yeast transcription factor SMP1. - *Arabidopsis thaliana* agamous protein (AG) [3], a probable transcription factor involved in regulating genes that determines stamen and carpel development in wild-type flowers. Mutations in the AG gene result in the replacement of the stamens by petals and the carpels by a new flower. - *Arabidopsis thaliana* homeotic proteins Apetala1 (AP1), Apetala3 (AP3) and Pistillata (PI) which act locally to specify the identity of
10 the floral meristem and to determine sepal and petal development [4]. - *Antirrhinum majus* and tobacco homeotic protein *deficiens* (DEFA) and *globosa* (GLO) [5]. Both proteins are transcription factors involved in the genetic control of flower development. Mutations in DEFA or GLO cause the transformation of petals into sepals and of stamens into carpels. - *Arabidopsis thaliana* putative transcription factors AGL1 to AGL6 [6]. - *Antirrhinum majus* morphogenetic protein DEF H33 (*squamosa*). In SRF, the conserved domain has been shown
15 [1] to be involved in DNA-binding and dimerization. A pattern that spans the complete length of the domain has been derived. The profile also spans the length of the MADS-box.

Consensus pattern: R-x-[RK]-x(5)-I-x-[DNGSK]-x(3)-[KR]-x(2)-T-[FY]-x-[RK](3)-x(2)-
[LIVM]-x-K(2)-A-x-E-[LIVM]-[STA]-x-L-x(4)-[LIVM]-x-[LIVM](3)-x(6)-[LIVMF]-x(2)-
20 [FY]

[1] Norman C., Runswick M., Pollock R., Treisman R. Cell 55:989-1003(1988).[2]
Passmore S., Maine G.T., Elble R., Christ C., Tye B.-K. *J. Mol. Biol.* 204:593-606(1988).[3]
Yanofsky M., Ma H., Bowman J., Drews G., Feldmann K.A., Meyerowitz E.M. *Nature*
346:35-39(1990).[4] Goto K., Meyerowitz E.M. *Genes Dev.* 8:1548-1560(1994).[5]
25 Troebner W., Ramirez L., Motte P., Hue I., Huijser P., Loennig W.-E., Saedler H., Sommer
H., Schwartz-Sommer Z. *EMBO J.* 11:4693-4704(1992).[6] Ma H., Yanofsky M.F.,
Meyerowitz E.M. *Genes Dev.* 5:484-495(1991).[E1]

671. Transketolase signatures

30 Transketolase (EC 2.2.1.1) (TK) catalyzes the reversible transfer of a two-carbon ketol unit from xylulose 5-phosphate to an aldose receptor, such as ribose 5-phosphate, to form sedoheptulose 7-phosphate and glyceraldehyde 3-phosphate. This enzyme, together with

transaldolase, provides a link between the glycolytic and pentose-phosphate pathways. TK requires thiamin pyrophosphate as a cofactor. In most sources where TK has been purified, it is a homodimer of approximately 70 Kd subunits. TK sequences from a variety of eukaryotic and prokaryotic sources [1,2] show that the enzyme has been evolutionarily conserved. In the peroxisomes of methylotrophic yeast *Hansenula polymorpha*, there is a highly related enzyme, dihydroxy-acetone synthase (DHAS) (EC 2.2.1.3) (also known as formaldehyde transketolase), which exhibits a very unusual specificity by including formaldehyde amongst its substrates. 1-deoxyxylulose-5-phosphate synthase (DXP synthase) [3] is an enzyme so far found in bacteria (gene *dxs*) and plants (gene *CLA1*) which catalyzes the thiamin pyrophosphate-dependent acyloin condensation reaction between carbon atoms 2 and 3 of pyruvate and glyceraldehyde 3-phosphate to yield 1-deoxy-D- xylulose-5-phosphate (*dxp*), a precursor in the biosynthetic pathway to isoprenoids, thiamin (vitamin B1), and pyridoxol (vitamin B6). DXP synthase is evolutionary related to TK. Two regions of TK have been selected as signature patterns. The first, located in the N-terminal section, contains a histidine residue which appears to function in proton transfer during catalysis [4]. The second, located in the central section, contains conserved acidic residues that are part of the active cleft and may participate in substrate-binding [4].

Consensus pattern: R-x(3)-[LIVMTA]-[DENQSTHKF]-x(5,6)-[GSN]-G-H-[PLIVMF]-[GSTA]-x(2)-[LIMC]-[GS]

Consensus pattern: G-[DEQGS]-[DN]-G-[PAEQ]-[ST]-[HQ]-x-[PAGM]-[LIVMYAC]-[DEFYW]-x(2)-[STAP]-x(2)-[RGA]

[1] Abedinia M., Layfield R., Jones S.M., Nixon P.F., Mattick J.S. *Biochem. Biophys. Res. Commun.* 183:1159-1166(1992).[2] Fletcher T.S., Kwee I.L., Nakada T., Largman C., Martin B.M. *Biochemistry* 31:1892-1896(1992).[3] Sprenger G.A., Schorken U., Wiegert T., Grolle S., De Graaf A.A., Taylor S.V., Begley T.P., Bringer-Meyer S., Sahm H. *Proc. Natl. Acad. Sci. U.S.A.* 94:12857-12862(1997).[4] Lindqvist Y., Schneider G., Ermler U., Sundstroem M. *EMBO J.* 11:2373-2379(1992).

672. Transmembrane 4 family signature

Recently a number of eukaryotic cell surface antigens have been found to be evolutionary related [1,2,3]. The proteins known to belong to this family are listed below: - Mammalian antigen CD9 (MIC3); A protein involved in platelet activation and aggregation. - Mammalian

leukocyte antigen CD37, expressed on B lymphocytes. - Mammalian leukocyte antigen CD53 (OX-44), which may be involved in growth regulation in hematopoietic cells. - Mammalian lysosomal membrane protein CD63 (melanoma-associated antigen ME491; antigen AD1). - Mammalian antigen CD81 (cell surface protein TAPA-1), which may play an important role in the regulation of lymphoma cell growth. - Mammalian antigen CD82 (protein R2; antigen C33; Kangai 1 (KAI1)), which associates with CD4 or CD8 and delivers costimulatory signals for the TCR/CD3 pathway. - Mammalian antigen CD151 (SFA-1; platelet-endothelial tetraspan antigen 3 (PETA-3)). - Mammalian cell surface glycoprotein A15 (TALLA-1; MXS1). - Mammalian novel antigen 2 (NAG-2). - Human tumor-associated antigen CO-029. - *Schistosoma mansoni* and *japonicum* 23 Kd surface antigen (SM23 / SJ23). These proteins share the following characteristics: they all seem to be type III membrane proteins (type III proteins are integral membrane proteins that contain a N-terminal membrane-anchoring domain which is not cleaved during biosynthesis and which functions both as a translocation signal and as a membrane anchor); they also contain three additional transmembrane regions, at least seven conserved cysteines residues, and are of approximately the same size (218 to 284 residues). These proteins are collectively known as the 'transmembrane 4 super family' (TM4) because they span the plasma membrane four times. A schematic diagram of the domain structure of these proteins is shown below.

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
 -----+-----+-----+ | TMa | Extra | TM2 | Cyt | TM3 | Extracellular | TM4 | Cyt | +-----+
 +-----+-----C-----C-----+-----CC-----C-----C-----+-----C-----+ ***** Cyt : cytoplasmic domain. TMa : transmembrane anchor. TM2 to TM4: transmembrane regions 2 to 4. 'C' : conserved cysteine. '*' : position of the pattern.

A conserved region that includes two cysteines and seems to be located in a short cytoplasmic loop between two transmembrane domains has been selected as a signature for these proteins.

Consensus pattern: G-x(3)-[LIVMF]-x(2)-[GSA]-[LIVMF](2)-G-C-x-[GA]-[STA]-x(2)-[EG]-x(2)-[CWN]-[LIVM](2)

[1] Levy S., Nguyen V.Q., Andria M.L., Takahashi S. J. Biol. Chem. 266:14597-

14602(1991).[2] Tomlinson M.G., Williams A.F., Wright M.D. Eur. J. Immunol. 23:136-

40(1993).[3] Barclay A.N., Birkeland M.L., Brown M.H., Beyers A.D., Davis S.J., Somoza C., Williams A.F. The leucocyte antigen factbooks. Academic Press, London / San Diego, (1993).

673. Tryptophan synthase alpha chain signature

Tryptophan synthase catalyzes the last step in the biosynthesis of tryptophan: the conversion of indoleglycerol phosphate and serine, to tryptophan and glyceraldehyde 3-phosphate [1,2]. It has two functional domains: one for the aldol cleavage of indoleglycerol phosphate to indole and glyceraldehyde 3-phosphate and the other for the synthesis of tryptophan from indole and serine. In bacteria and plants [3], each domain is found on a separate subunit (alpha and beta chains), while in fungi the two domains are fused together on a single multifunctional protein. A conserved region that contains three conserved acidic residues has been selected as a signature pattern for the alpha chain. The first and the third acidic residues are believed to serve as proton donors/acceptors in the enzyme's catalytic mechanism.

Consensus pattern: [LIVM]-E-[LIVM]-G-x(2)-[FYC]-[ST]-[DE]-[PA]-[LIVMY]-[AGLI]-[DE]-G

[1] Crawford I.P. Annu. Rev. Microbiol. 43:567-600(1989).[2] Hyde C.C., Miles E.W. Bio/Technology 8:27-32(1990).[3] Berlyn M.B., Last R.L., Fink G.R. Proc. Natl. Acad. Sci. U.S.A. 86:4604-4608(1989).

674. Tryptophan synthase beta chain pyridoxal-phosphate attachment site

Tryptophan synthase catalyzes the last step in the biosynthesis of tryptophan: the conversion of indoleglycerol phosphate and serine, to tryptophan and glyceraldehyde 3-phosphate [1,2]. It has two functional domains: one for the aldol cleavage of indoleglycerol phosphate to indole and glyceraldehyde 3-phosphate and the other for the synthesis of tryptophan from indole and serine. In bacteria and plants [3], each domain is found on a separate subunit (alpha and beta chains), while in fungi the two domains are fused together on a single multifunctional protein. The beta chain of the enzyme requires pyridoxal-phosphate as a cofactor. The pyridoxal-phosphate group is attached to a lysine residue. The region around this lysine residue also contains two histidine residues which are part of the pyridoxal-phosphate binding site. The signature pattern for the tryptophan synthase beta chain is derived from that conserved region.

-Consensus pattern: [LIVM]-x-H-x-G-[STA]-H-K-x-N [K is the pyridoxal-P attachment site]
[1] Crawford I.P. Annu. Rev. Microbiol. 43:567-600(1989).[2] Hyde C.C., Miles E.W. Bio/Technology 8:27-32(1990).[3] Berlyn M.B., Last R.L., Fink G.R. Proc. Natl. Acad. Sci. U.S.A. 86:4604-4608(1989).

675. Serine proteases, trypsin family, active sites

The catalytic activity of the serine proteases from the trypsin family is provided by a charge relay system involving an aspartic acid residue hydrogen-bonded to a histidine, which itself is hydrogen-bonded to a serine. The sequences in the vicinity of the active site serine and histidine residues are well conserved in this family of proteases [1]. A partial list of proteases known to belong to the trypsin family is shown below. - Acrosin. - Blood coagulation factors VII, IX, X, XI and XII, thrombin, plasminogen, and protein C. - Cathepsin G. -

Chymotrypsins. - Complement components C1r, C1s, C2, and complement factors B, D and I. - Complement-activating component of RA-reactive factor. - Cytotoxic cell proteases (granzymes A to H). - Duodenase I. - Elastases 1, 2, 3A, 3B (protease E), leukocyte (medullasin). - Enterokinase (EC 3.4.21.9) (enteropeptidase). - Hepatocyte growth factor activator. - Hepsin. - Glandular (tissue) kallikreins (including EGF-binding protein types A, B, and C, NGF-gamma chain, gamma-renin, prostate specific antigen (PSA) and tonin). - Plasma kallikrein. - Mast cell proteases (MCP) 1 (chymase) to 8. - Myeloblastin (proteinase 3) (Wegener's autoantigen). - Plasminogen activators (urokinase-type, and tissue-type). - Trypsins I, II, III, and IV. - Trypsases. - Snake venom proteases such as ancrod, batroxobin, cerastobin, flavoxobin, and protein C activator. - Collagenase from common cattle grub and collagenolytic protease from Atlantic sand fiddler crab. - Apolipoprotein(a). - Blood fluke cercarial protease. - Drosophila trypsin like proteases: alpha, easter, snake-locus. - Drosophila protease stubble (gene sb). - Major mite fecal allergen Der p III. All the above proteins belong to family S1 in the classification of peptidases[2,E1] and originate from eukaryotic species. It should be noted that bacterial proteases that belong to family S2A are similar

enough in the regions of the active site residues that they can be picked up by the same patterns. These proteases are listed below. - Achromobacter lyticus protease I. - Lysobacter alpha-lytic protease. - Streptogrisin A and B (Streptomyces proteases A and B). - Streptomyces griseus glutamyl endopeptidase II. - Streptomyces fradiae proteases 1 and 2. Consensus pattern: [LIVM]-[ST]-A-[STAG]-H-C [H is the active site residue]

Consensus pattern: [DNSTAGC]-[GSTAPIMVQH]-x(2)-G-[DE]-S-G-[GS]-[SAPHV]-[LIVMFYWH]-[LIVMFYSTANQH] [S is the active site residue]

[1] Brenner S. Nature 334:528-530(1988).[2] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).[E1]

676. (tsp) Thrombospondin type 1 domain

5 [1] Bork P; FEBS lett 1993;327:125-130.

677. Tubulin subunits alpha, beta, and gamma signature

10 Tubulins [1,2], the major constituent of microtubules are dimeric proteins which consist of two closely related subunits (alpha and beta). Tubulin binds two molecules of GTP at two different sites (N and E). At the E (Exchangeable) site, GTP is hydrolyzed during incorporation into the microtubule. Near the E site is an invariant region rich in glycines which is found in both chains and which is now [3] said to control the access of the nucleotide to its binding site. A signature pattern was developed from this region. With the exception of the simple eukaryotes, most species express a variety of closely related alpha and beta isotypes. In most species there is a third member of the tubulin family: gamma tubulin. Gamma tubulin is found at microtubule organizing centers (MTOC) such as the spindle poles or the centrosome, suggesting that it is involved in the minus-end nucleation of microtubule assembly [4].

15 20 Consensus pattern: [SAG]-G-G-T-G-[SA]-G

[1] Cleveland D.W., Sullivan K.F. Annu. Rev. Biochem. 54:331-365(1985).[2] Joshi H.C., Cleveland D.W. Cell Motil. Cytoskeleton 16:159-163(1990).[3] Hesse J., Thierauf M., Ponstingl H. J. Biol. Chem. 262:15472-15475(1987).[4] Joshi H.C. BioEssays 15:637-643(1993).

25

Tubulin-beta mRNA autoregulation signal

The stability of beta-tubulin mRNAs are autoregulated by their own translation product [1]. Unpolymerized tubulin subunits bind directly (or activate a factor(s) which binds co-translationally) to the nascent N-terminus of beta-tubulin. This binding is transduced through the adjacent ribosomes to activate an RNase that degrades the polysome-bound mRNA. The recognition element has been shown to be the first four amino acids of beta-tubulin: Met-Arg-Glu-Ile. Mutations to this sequence abolish the autoregulation effect (except for the

30

replacement of Glu by Asp); transposition of this sequence to an internal region of a polypeptide also suppresses the autoregulatory effect.

Consensus pattern: <M-R-[DE]-[IL]

[1] Cleveland D.W. Trends Biochem. Sci. 13:339-343(1988).

5

678. (tRNA-synt 2c) Aminoacyl-transfer RNA synthetases class-II signatures. Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure. The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is different to the Rossmann fold observed for the class I synthetases [7]. Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns have been derived from two of these regions.

Consensus pattern: [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]-

Consensus pattern: [GSTALVF]-{DENQHRKP}-[GSTA]-[LIVMF]-[DE]-R-[LIVMF]-x-[LIVMSTAG]-[LIVMFY]-

25

[1] Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).[2] Delarue M., Moras D. BioEssays 15:675-687(1993).[3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).[4] Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991). [5] Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).[6] Cusack S. Biochimie 75:1077-1081(1993).[7] Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-255(1990).[8] Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).

30

679. UBA-domain

The UBA-domain (ubiquitin associated domain) is a novel sequence motif found in several proteins having connections to ubiquitin and the ubiquitination pathway. The structure of the UBA domain consists of a compact three helix bundle [1]. Number of members: 84

[1] Structure of a human DNA repair protein UBA domain that interacts with HIV-1 Vpr. Dieckmann T, Withers-Ward ES, Jarosinski MA, Liu CF, Chen IS, Feigon J; Nat Struct Biol 1998;5:1042-1047.

680. UBX domain

Domain present in ubiquitin-regulatory proteins. Present in FAF1 and Shp1p. Number of members: 19

[1] The UBA domain: a sequence motif present in multiple enzyme classes of the ubiquitination pathway. Hofmann K, Bucher P; Trends Biochem Sci 1996;21:172-173.

681. (UCH) Ubiquitin carboxyl-terminal hydrolases family 1 cysteine active site

Ubiquitin carboxyl-terminal hydrolases (UCH) (deubiquitinating enzymes) [1,2] are thiol proteases that recognize and hydrolyze the peptide bond at the C-terminal glycine of ubiquitin. These enzymes are involved in the processing of poly-ubiquitin precursors as well as that of ubiquitinated proteins. There are two distinct families of UCH. The first class consist of enzymes of about 25 Kd and is currently represented by: - Mammalian isozymes L1 and L3. - Yeast YUH1. - Drosophila Uch. One of the active site residues of class-I UCH [3] is a cysteine. A signature pattern has been derived from the region around that residue. Consensus pattern: Q-x(3)-N-[SA]-C-G-x(3)-[LIVM](2)-H-[SA]-[LIVM]-[SA] [C is the active site residue

[1] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991).[2]

D'andrea A., Pellman D. Crit. Rev. Biochem. Mol. Biol. 33:337-352(1998).[3] Johnston S.C., Larsen C.N., Cook W.J., Wilkinson K.D., Hill C.P. EMBO J. 16:3787-3796(1997).[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:461-486(1994).

682. Ubiquitin carboxyl-terminal hydrolases family 2 signatures (UCH-1)

Ubiquitin carboxyl-terminal hydrolases (UCH) (deubiquitinating enzymes) [1,2] are thiol proteases that recognize and hydrolyze the peptide bond at the C-terminal glycine of ubiquitin. These enzymes are involved in the processing of poly-ubiquitin precursors as well as that of ubiquitinated proteins. There are two distinct families of UCH. The second class consist of largeproteins (800 to 2000 residues) and is currently represented by: - Yeast UBP1, UBP2, UBP3, UBP4 (or DOA4/SSV7), UBP5, UBP7, UBP9, UBP10, UBP11, UBP12, UBP13, UBP14, UBP15 and UBP16. - Human tre-2. - Human isopeptidase T. - Human isopeptidase T-3. - Mammalian Ode-1. - Mammalian Unp. - Mouse Dub-1. - Drosophila fat facets protein (gene *faf*). - Mammalian *faf* homolog. - Drosophila D-Ubp-64E. - *Caenorhabditis elegans* hypothetical protein R10E11.3. - *Caenorhabditis elegans* hypothetical protein K02C4.3. These proteins only share two regions of similarity. The first region contains a conserved cysteine which is probably implicated in the catalytic mechanism. The second region contains two conserved histidines residues, one of which is also probably implicated in the catalytic mechanism. Signature patterns for both conserved regions have been developed.

Consensus pattern: G-[LIVMFY]-x(1,3)-[AGC]-[NASM]-x-C-[FYW]-[LIVMC]-[NST]-[SACV]-x-[LIVMS]-Q [C is the putative active site residue]

Consensus pattern: Y-x-L-x-[SAG]-[LIVMFT]-x(2)-H-x-G-x(4,5)-G-H-Y [The two H's are putative active site residues]

[1] Jentsch S., Seufert W., Hauser H.-P. *Biochim. Biophys. Acta* 1089:127-139(1991). [2] D'andrea A., Pellman D. *Crit. Rev. Biochem. Mol. Biol.* 33:337-352(1998). [3] Rawlings N.D., Barrett A.J. *Meth. Enzymol.* 244:461-486(1994).

683. Ubiquitin carboxyl-terminal hydrolases family 2 signatures (UCH-2)

Ubiquitin carboxyl-terminal hydrolases (UCH) (deubiquitinating enzymes) [1,2] are thiol proteases that recognize and hydrolyze the peptide bond at the C-terminal glycine of ubiquitin. These enzymes are involved in the processing of poly-ubiquitin precursors as well as that of ubiquitinated proteins. There are two distinct families of UCH. The second class consist of largeproteins (800 to 2000 residues) and is currently represented by: - Yeast UBP1, UBP2, UBP3, UBP4 (or DOA4/SSV7), UBP5, UBP7, UBP9, UBP10, UBP11, UBP12,

UBP13, UBP14, UBP15 and UBP16. - Human tre-2. - Human isopeptidase T. - Human isopeptidase T-3. - Mammalian Ode-1. - Mammalian Unp. - Mouse Dub-1. - Drosophila fat facets protein (gene faf). - Mammalian faf homolog. - Drosophila D-Ubp-64E. -

Caenorhabditis elegans hypothetical protein R10E11.3. - Caenorhabditis elegans hypothetical protein K02C4.3. These proteins only share two regions of similarity. The first region contains a conserved cysteine which is probably implicated in the catalytic mechanism. The second region contains two conserved histidines residues, one of which is also probably implicated in the catalytic mechanism. Signature patterns for both conserved regions have been developed.

Consensus pattern: G-[LIVMFY]-x(1,3)-[AGC]-[NASM]-x-C-[FYW]-[LIVMC]-[NST]-[SACV]-x-[LIVMS]-Q [C is the putative active site residue]

Consensus pattern: Y-x-L-x-[SAG]-[LIVMFT]-x(2)-H-x-G-x(4,5)-G-H-Y [The two H's are putative active site residues]

[1] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991).[2] D'andrea A., Pellman D. Crit. Rev. Biochem. Mol. Biol. 33:337-352(1998).[3] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:461-486(1994).

684. UDP-glycosyltransferases signature

UDP glycosyltransferases (UGT) are a superfamily of enzymes that catalyzes the addition of the glycosyl group from a UTP-sugar to a small hydrophobic molecule. This family currently consist of: - Mammalian UDP-glucuronosyl transferases (UDPGT) [1,2]. A large family of membrane-bound microsomal enzymes which catalyze the transfer of glucuronic acid to a wide variety of exogenous and endogenous lipophilic substrates. These enzymes are of major importance in the detoxification and subsequent elimination of xenobiotics such as drugs and carcinogens. - A large number of putative UDPGT from Caenorhabditis elegans. -

Mammalian 2-hydroxyacylsphingosine 1-beta-galactosyltransferase [3] (also known as UDP-galactose-ceramide galactosyltransferase). This enzyme catalyzes the transfer of galactose to ceramide, a key enzymatic step in the biosynthesis of galactocerebrosides, which are abundant sphingolipids of the myelin membrane of the central nervous system and peripheral nervous system. - Plants flavonol O(3)-glucosyltransferase. An enzyme [4] that catalyzes the transfer of glucose from UDP-glucose to a flavanol. This reaction is essential and one of the last steps in anthocyanin pigment biosynthesis. - Baculoviruses ecdysteroid UDP-

glucosyltransferase (EC 2.4.1.-) [5] (egt). This enzyme catalyzes the transfer of glucose from UDP-glucose to ectysteroids which are insect molting hormones. The expression of egt in the insect host interferes with the normal insect development by blocking the molting process. - Prokaryotic zeaxanthin glucosyl transferase (gene crtX), an enzyme involved in carotenoid biosynthesis and that catalyses the glycosylation reaction which converts zeaxanthin to zeaxanthin-beta- diglucoside. - Streptomyces macrolide glycosyltransferases [6]. These enzymes specifically inactivates macrolide antibiotics via 2'-O-glycosylation using UDP-glucose. These enzymes share a conserved domain of about 50 amino acid residues located in their C-terminal section and from which a pattern has been extracted to detect them.

Consensus pattern: [FW]-x(2)-Q-x(2)-[LIVMYA]-[LIMV]-x(4,6)-[LVGAC]-[LVFYA]-[LIVMF]-[STAGCM]-[HNQ]-[STAGC]-G-x(2)-[STAG]-x(3)-[STAGL]-[LIVMFA]-x(4)-[PQR]-[LIVMT]-x(3)-[PA]-x(3)-[DES]-[QEHN]

[1] Dutton G.J. (In) Glucuronidation of drugs and other compounds, Dutton G.J., Ed., pp 1-78, CRC Press, Boca Raton, (1980).[2] Burchell B., Nebert D.W., Nelson D.R., Bock K.W., Iyanagi T., Jansen P.L., Lancet D., Mulder G.J., Chowdhury J.R., Siest G., Tephly T.R., Mackenzie P.I. DNA Cell Biol. 10:487-494(1991).[3] Schulte S., Stoffel W. Proc. Natl. Acad. Sci. U.S.A. 90:10265-10269(1993).[4] Furtak D., Schiefelbein J.W., Johnston F., Nelson O.E. Jr. Plant Mol. Biol. 11:473-481(1988).[5] O'Reilly D.R., Miller L.K. Science 245:1110-1112(1989).[6] Hernandez C., Olano C., Mendez C., Salas J.A. Gene 134:139-140(1993).

685. UDP-glucose/GDP-mannose dehydrogenase family

The UDP-glucose/GDP-mannose dehydrogenases are a small group of enzymes which possesses the ability to catalyze the NAD-dependent 2-fold oxidation of an alcohol to an acid without the release of an aldehyde intermediate [2]. Number of members: 55

[1] Purification and characterization of guanosine diphospho-D-mannose dehydrogenase. A key enzyme in the biosynthesis of alginate by *Pseudomonas aeruginosa*. Roychoudhury S, May TB, Gill JF, Singh SK, Feingold DS, Chakrabarty AM; J Biol Chem 1989;264:9380-9385. [2] Properties and kinetic analysis of UDP-glucose dehydrogenase from group A streptococci. Irreversible inhibition by UDP-chloroacetol. Campbell RE, Sala RF, van de Rijn I, Tanner ME; J Biol Chem 1997;272:3416-3422.

686. Uracil-DNA glycosylase signature

Uracil-DNA glycosylase (EC 3.2.2.-) (UNG) [1] is a DNA repair enzyme that excises uracil residues from DNA by cleaving the N-glycosylic bond. Uracil in DNA can arise as a result of misincorporation of dUMP residues by DNA polymerase or deamination of cytosine. The sequence of uracil-DNA glycosylase is extremely well conserved [2] in bacteria and eukaryotes as well as in herpes viruses. More distantly related uracil-DNA glycosylases are also found in poxviruses [3]. In eukaryotic cells, UNG activity is found in both the nucleus and the mitochondria. Human UNG1 protein is transported to both the mitochondria and the nucleus [4]. The N-terminal 77 amino acids of UNG1 seem to be required for mitochondrial localization [4], but the presence of a mitochondrial transitpeptide has not been directly demonstrated. As a signature for this type of enzyme, the most N-terminal conserved region has been selected. This region contains an aspartic acid residue which has been proposed, based on X-ray structures [5,6] to act as a general base in the catalytic mechanism.

Consensus pattern: [KR]-[LIV]-[LIVC]-[LIVM]-x-G-[QI]-D-P-Y [D is the active site residue]-

[1] Sancar A., Sancar G.B. *Annu. Rev. Biochem.* 57:29-67(1988).[2] Olsen L.C., Aasland R., Wittwer C.U., Krokan H.E., Helland D.E. *EMBO J.* 8:3121-3125 (1989).[3] Upton C., Stuart D.T., McFadden G. *Proc. Natl. Acad. Sci. U.S.A.* 90:4518-4522(1993).[4] Slupphaug G., Markussen F.-H., Olsen L.C., Aasland R., Aarsaether N., Bakke O., Krokan H.E., Helland D.E. *Nucleic Acids Res.* 21:2579-2584(1993).[5] Savva R., McAuley-Hecht K., Brown T., Pearl L. *Nature* 373:487-493(1995).[6] Mol C.D., Arvai A.S., Slupphaug G., Kavli B., Alseth I., Krokan H.E., Tainer J.A. *Cell* 80:869-878(1995).[7] Muller S.J., Caradonna S. *Biochim. Biophys. Acta* 1088:197-207(1991).[8] Meyer-Siegler K., Mauro D.J., Seal G., Wurzer J., Deriel J.K., Sirover M.A. *Proc. Natl. Acad. Sci. U.S.A.* 88:8460-8464(1991).[9] Muller S.J., Caradonna S. *J. Biol. Chem.* 268:1310-1319(1993).[10] Barnes D.E., Lindahl T., Sedgwick B. *Curr. Opin. Cell Biol.* 5:424-433(1993).

687. Uncharacterized protein family UPF0001 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Yeast chromosome II hypothetical protein YBL036c. - *Caenorhabditis elegans* hypothetical protein F09E5.8. - *Bacillus subtilis* hypothetical protein ylmE. - *Escherichia coli* hypothetical

protein yggS and HI0090, the corresponding *Haemophilus influenzae* protein. - *Helicobacter pylori* hypothetical protein HP0395. - *Mycobacterium tuberculosis* hypothetical protein MtCY270.20. - *Synechocystis* strain PCC 6803 hypothetical protein slr0556. - A

Pseudomonas aeruginosa hypothetical protein in pilT 5'region. - A *Vibrio alginolyticus*

- 5 hypothetical protein in pilT 5'region. These are proteins of from 25 to 30 Kd which contain a number of conserved regions. The best conserved region which is located in the first third of these proteins has been selected as a signature pattern.

Consensus pattern: [FW]-H-[FM]-[IV]-G-x-[LIV]-Q-x-[NKR]-K-x(3)-[LIV]

[1] Bairoch A., Rudd K.E. Unpublished observations (1996).

10

688. Uncharacterized protein family UPF0003 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -

Escherichia coli protein aefA. - *Escherichia coli* hypothetical protein yggB. - *Escherichia coli* hypothetical protein yjeP and HI0195.1, the corresponding *Haemophilus influenzae* protein. -

Escherichia coli hypothetical protein ynaI. - *Bacillus subtilis* hypothetical protein yhdY. -

Helicobacter pylori hypothetical protein HP0415. - *Synechocystis* strain PCC 6803

hypothetical protein slr0639. - *Archaeoglobus fulgidus* hypothetical protein AF1546. -

Methanococcus jannaschii hypothetical protein MJ0170. - *Methanococcus jannaschii*

hypothetical protein MJ1143. The size of these proteins range from 30 to 120 Kd. They all contain a number of transmembrane regions. The best conserved region which is located in and just after the last potential transmembrane region has been selected as a signature pattern,.

Consensus pattern: G-[STIF]-V-x(2)-[LIVM]-x(6)-[LIVMF]-x(3)-[DQ]-x(3)-[LIV]- x-[LIV]-

25 P-N-x(2)-[LIVMF]-[LIVFSTA]-x(5)-N

[1] Bairoch A. Unpublished observations (1997).

689. Uncharacterized protein family UPF0004 signature

- 30 The following uncharacterized proteins have been shown [1] to share regions of similarities: -

Escherichia coli hypothetical protein yliG. - *Escherichia coli* hypothetical protein yleA and

HI0019, the corresponding *Haemophilus influenzae* protein. - *Bacillus subtilis* hypothetical

protein yqeV. - *Helicobacter pylori* hypothetical protein HP0269. - *Helicobacter pylori*

559

hypothetical protein HP0285. - Mycoplasma iowae hypothetical protein in 16S RNA
 5'region. - Mycobacterium leprae hypothetical protein B2235_C2_195. - Pseudomonas
 aeruginosa hypothetical protein in hemL 3'region. - Synechocystis strain PCC 6803
 hypothetical protein slr0082. - Synechocystis strain PCC 6803 hypothetical protein slI0996. -
 5 Methanococcus jannaschii hypothetical protein MJ0865. - Methanococcus jannaschii
 hypothetical protein MJ0867. - Caenorhabditis elegans hypothetical protein F25B5.5. The size
 of these proteins range from 47 to 61 Kd. They contain six conserved cysteines, three of
 which are clustered in a region that can be used as a signature pattern.

Consensus pattern: [LIVM]-x-[LIVMT]-x(2)-G-C-x(3)-C-[STAN]-[FY]-C-x-[LIVM]-x(4)-

10 G

[1] Bairoch A. Unpublished observations (1997).

690. Uncharacterized protein family UPF0005 signature

The following proteins seem to be evolutionarily related [1]: - Mammalian protein TEGT
 (Testis Enhanced Gene Transcript). - Escherichia coli hypothetical protein yccA and HI0044,
 the corresponding Haemophilus influenzae protein. - A probable Pseudomonas aeruginosa
 ortholog of yccA. These are proteins of about 25 Kd which seem to contain seven
 transmembrane domains. A signature pattern that corresponds to a region that starts with the
 beginning of the third transmembrane domain and ends in the middle of the fourth one has
 been developed.

Consensus pattern: G-[LIVM](2)-[SA]-x(5,8)-G-x(2)-[LIVM]-G-P-x-L-x(4)-[SAG]-x(4,6)-
 [LIVM](2)-x(2)-A-x(3)-T-A-[LIVM](2)-F

[1] Walter L., Marynen P., Szpirer J., Levan G., Guenther E. Genomics 28:301-304(1995).

691. Uncharacterized protein family UPF0006 signatures

The following uncharacterized proteins have been shown [1] to share regions of similarities: -
 Yeast chromosome II hypothetical protein YBL055c. - Escherichia coli hypothetical protein
 ycfH and HI0454, the corresponding Haemophilus influenzae protein. - Escherichia coli
 hypothetical protein yigW. - Escherichia coli hypothetical protein yjjV and HI0081, the
 corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein yabD.
 - Haemophilus influenzae hypothetical protein HI1664. - Mycoplasma genitalium

hypothetical protein MG009. These are proteins of from 24 to 47 Kd which contain a number of conserved regions. They can be picked up in the database by the following patterns.

Consensus pattern: [LIVMFY](2)-D-[STA]-H-x-H-[LIVMF]-[DN

Consensus pattern: P-[LIVM]-x-[LIVM]-H-x-R-x-[TA]-x-[DE

5 Consensus pattern: [LVSA]-[LIVA]-x(2)-[LIVM]-[PS]-x(3)-L-[LIVM]-[LIVMS]-E-T- D-x-P

[1] Bairoch A., Rudd K.E. Unpublished observations (1995).

10 692. Uncharacterized protein family UPF0007 signature

The following proteins seems to be evolutionary related [1]: - Escherichia coli hypothetical protein ygbP and HI0672, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein yacM. - Mycobacterium tuberculosis hypothetical protein MtCY06G11.29c. - Synechocystis strain PCC 6803 hypothetical protein slr0951. - A Rhodobacter capsulatus hypothetical protein in nifR3 5'region. Except for the Rhodobacter protein which contains a C-terminal extension, all these proteins have from 225 to 236 amino acids. They are hydrophilic proteins that can be picked up in the database by the following pattern.

Consensus pattern: V-L-[IV]-H-D-[GA]-A-R

20 [1] Bairoch A. Unpublished observations (1997).

693. Uncharacterized protein family UPF0015 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -

25 Yeast chromosome II hypothetical protein YBR002c. - Yeast chromosome XIII hypothetical protein YMR101c. - Escherichia coli hypothetical protein yaeU and HI0920, the corresponding Haemophilus influenzae protein. - Helicobacter pylori hypothetical protein HP1221. - Mycobacterium leprae hypothetical protein B1937_F2_65. - A Corynebacterium glutamicum hypothetical protein in aroF 3'region. - A Streptomyces fradiae hypothetical
30 protein in transposon Tn4556. - Synechocystis strain PCC 6803 hypothetical protein sll0505. - Methanococcus jannaschii hypothetical protein MJ1372. These are proteins of about 26 to 40 Kd whose central region is well conserved. They can be picked up in the database by the following pattern.

Consensus pattern: [DE]-[LIVMF](3)-R-T-[SG]-G-x(2)-R-x-S-x-[FY]-[LIVM](2)-W-Q-
[1] Wolfe K.H., Lohan A.J.E. Yeast 10:S41-S46(1994).

5 694. Uncharacterized protein family UPF0016 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -
Yeast hypothetical protein YBR187w. - Fission yeast hypothetical protein SpAC17G8.08c. -
Mouse protein pFT27. - Synechocystis strain PCC 6803 hypothetical protein sll0615. These
are hydrophobic proteins of 200 to 320 amino acids that seem to contain six or seven
10 transmembrane domains. A conserved region which seems, in the eukaryotic proteins of this
family, to directly follow the second transmembrane domain has been selected as a signature
pattern.

Consensus pattern: E-[LIVM]-G-D-K-T-F-[LIVMF](2)-A-
[1] Bairoch A. Unpublished observations (1996).

15 695. Uncharacterized protein family UPF0021 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -
Yeast chromosome VII hypothetical protein YGL211w. - Dictyostelium discoideum protein
veg136. - Methanococcus jannaschii hypothetical proteins MJ1157 and MJ1478. These are
20 proteins of from 300 to 360 residues. They can be picked up in the database by the following
pattern which is located in their N-terminal section.

Consensus pattern: C-K-x(2)-F-x(4)-E-x(22,23)-S-G-G-K-D
[1] Bairoch A. Unpublished observations (1997).

25 696. Uncharacterized protein family UPF0023 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -
Mouse protein 22A3. - Yeast chromosome XII hypothetical protein YLR022c. -
30 Caenorhabditis elegans hypothetical protein W06E11.4. - Methanococcus jannaschii
hypothetical protein MJ0592. These are hydrophilic proteins of about 30 Kd. They can be
picked up in the database by the following pattern.

Consensus pattern: D-x-D-E-[LIV]-L-x(4)-V-F-x(3)-S-K-G-

[1] Bairoch A. Unpublished observations (1997).

697. Uncharacterized protein family UPF0024 signature. The following uncharacterized
 5 proteins have been shown [1] to share regions of similarities: - Escherichia coli hypothetical
 protein ygbO and HI0701, the corresponding Haemophilus influenzae protein. - Helicobacter
 pylori hypothetical protein HP0926. - Yeast chromosome XV hypothetical protein YOR243c.
 - Caenorhabditis elegans hypothetical protein B0024.11. - Methanococcus jannaschii
 hypothetical proteins MJ0588 and MJ1364. These are hydrophilic proteins of from 39 to 77
 10 Kd. They can be picked up in the database by the following pattern.

Consensus pattern: G-x-K-D-[KR]-x-A-[LV]-T-x-Q-x-[LIVF]-[SGC]-

[1] Bairoch A. Unpublished observations (1997).

698. Uncharacterized protein family UPF0025 signature
 The following uncharacterized proteins have been shown [1] to share regions of similarities: -
 Escherichia coli hypothetical protein yfcE. - Bacillus subtilis hypothetical protein ysnB. -
 20 Mycoplasma genitalium and pneumoniae hypothetical protein MG207. - Methanococcus
 jannaschii hypothetical proteins MJ0623 and MJ0936. These are hydrophilic proteins of
 about 20 Kd. They can be picked up in the database by the following pattern.

Consensus pattern: D-V-[LIV]-x(2)-G-H-[ST]-H-x(12)-[LIVMF]-N-P-G

[1] Bairoch A. Unpublished observations (1997).

699. Uncharacterized protein family UPF0029 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -
 Yeast chromosome III hypothetical protein YCR59c. - Yeast chromosome IV hypothetical
 30 protein YDL177C. - Escherichia coli hypothetical protein yigZ and HI0722, the
 corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein yvyE.
 - A Thermus aquaticus hypothetical protein in pol 5' region. These proteins can be picked up
 in the database by the following pattern.

563

Consensus pattern: G-x(2)-[LIVM](2)-x(2)-[LIVM]-x(4)-[LIVM]-x(5)-[LIVM](2)-x- R-
[FYW](2)-G-G-x(2)-[LIVM]-G

[1] Koonin E.V., Bork P., Sander C. EMBO J. 13:493-503(1994).

5

700. Uncharacterized protein family UPF0030 signature

The following uncharacterized proteins have been shown [1] to be highly similar: - Yeast
chromosome VI hypothetical protein YFL060c. - Yeast chromosome XIII hypothetical
protein YMR095c. - Yeast chromosome XIV hypothetical protein YNL334c. - Bacillus
subtilis hypothetical protein yaaE. - Haemophilus influenzae hypothetical protein HI1648. -
Methanococcus jannaschii hypothetical protein MJ1661. These are hydrophilic proteins of
about 19 to 25 Kd. They can be picked up in the database by the following pattern.

Consensus pattern: [GA]-L-I-[LIV]-P-G-G-E-S-T-[STA]

[1] Bairoch A. Unpublished observations (1997).

15

701. Uncharacterized protein family UPF0032 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -
Escherichia coli hypothetical protein yigU and HI0188, the corresponding Haemophilus
influenzae protein. - Bacillus subtilis hypothetical protein ycbT. - Mycobacterium
tuberculosis hypothetical protein MtCY49.33c and U2126A, the corresponding
Mycobacterium leprae protein. - Synechocystis strain PCC 6803 hypothetical protein sll0194.
- Odontella sinensis and Porphyra purpurea chloroplast hypothetical protein ycf43. These
proteins have from 245 to 317 amino acids and seem to contain at least six or seven
transmembrane regions. A conserved region located in the central section of these proteins
has been developed as a signature pattern,.

Consensus pattern: Y-x(2)-F-[LIVMA](2)-x-L-x(4)-G-x(2)-F-[EQ]-[LIVMF]-P- [LIVM] -

[1] Bairoch A., Rudd K.E. Unpublished observations (1996).

30

702. Uncharacterized protein family UPF0034 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: -
Escherichia coli hypothetical protein yhdG and HI0979, the corresponding Haemophilus

influenzae protein. - Escherichia coli hypothetical protein yjbN and HI0634, the corresponding Haemophilus influenzae protein. - Escherichia coli hypothetical protein yohI and HI0270, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein yacF. - Rhodobacter capsulatus protein nifR3 and related proteins in Azospirillum brasilense and Rhizobium leguminosarum. - Synechocystis strain PCC 6803 hypothetical protein slr0644. - Synechocystis strain PCC 6803 hypothetical protein slr0926. - Caenorhabditis elegans hypothetical protein C45G9.2. - Yeast protein SMM1. - Yeast hypothetical protein YLR401c. - Yeast hypothetical protein YLR405w. - Yeast hypothetical protein YML080w. Although it has been proposed [2] that Rhodobacter capsulatus nifR3 is a transcriptional regulatory protein, it is believed that these proteins constitute a family of enzymes whose active site could include a conserved cysteine which has been used as the central part of a signature pattern.

Consensus pattern: [LIVM]-[DNG]-[LIVM]-N-x-G-C-P-x(3)-[LIVMASQ]-x(5)-G-[SAC]
[1] Bairoch A., Rudd K.E. Unpublished observations (1995).[2] Foster-Hartnett D., Cullen P.J., Gabbert K.K., Kranz R.G. Mol. Microbiol. 8:903-914(1993).

703. Uncharacterized protein family UPF0038 signature

The following uncharacterized proteins have been shown [1] to share regions of similarities: - Escherichia coli hypothetical protein yacE and HI0890, the corresponding Haemophilus influenzae protein. - Mycobacterium tuberculosis hypothetical protein MtCY01B2.23 and O410, the corresponding Mycobacterium leprae protein. - Synechocystis strain PCC 6803 hypothetical protein slr0553. - Other hypothetical proteins from Aeromonas hydrophila, Bacteroides nodosus, Neisseria gonorrhoeae, Pseudomonas putida, Thermus thermophilus and Xanthomonas campestris. - Human hypothetical protein pOV-2. - Yeast hypothetical protein YDR196C. - Caenorhabditis elegans hypothetical protein T05G5.5. These proteins all contain, in their N-terminal extremity, an ATP/GTP-binding motif 'A' (P-loop) (see <PDOC00017>). The size of these proteins range from 200 to 290 residues (with the exception of the Mycobacterial sequences which are 410 residues long). A conserved region some 50 residues away from the ATP-binding P-loop has been developed as a signature pattern.

Consensus pattern: G-x-[LI]-x-R-x(2)-L-x(4)-F-x(8)-[LIV]-x(5)-P-x-[LIV] -
[1] Rudd K.E., Bairoch A. Unpublished observations (1997).

704. Ubiquitin-conjugating enzymes active site

Ubiquitin-conjugating enzymes (UBC or E2 enzymes) [1,2,3] catalyze the covalent attachment of ubiquitin to target proteins. An activated ubiquitin moiety is transferred from an ubiquitin-activating enzyme (E1) to E2 which later ligates ubiquitin directly to substrate proteins with or without the assistance of 'N-end' recognizing proteins (E3). In most species there are many forms of UBC (at least 9 in yeast) which are implicated in diverse cellular functions. A cysteine residue is required for ubiquitin-thiolester formation. There is a single conserved cysteine in UBC's and the region around that residue is conserved in the sequence of known UBC isozymes. That region has been used as a signature pattern.

Consensus pattern: [FYWLSP]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x-[LIV]-C-[LIV]-x-[LIV] [C is the active site residue]

[1] Jentsch S., Seufert W., Sommer T., Reins H.-A. Trends Biochem. Sci. 15:195-198(1990). [2] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991). [3] Hershko A. Trends Biochem. Sci. 16:265-268(1991).

705. Uroporphyrinogen decarboxylase signatures

Uroporphyrinogen decarboxylase (URO-D), the fifth enzyme of the heme biosynthetic pathway, catalyzes the sequential decarboxylation of the four acetyl side chains of uroporphyrinogen to yield coproporphyrinogen [1]. URO-D deficiency is responsible for the Human genetic diseases familial porphyria cutanea tarda (fPCT) and hepatoerythropoietic porphyria (HEP). The sequence of URO-D has been well conserved throughout evolution.

The best conserved region is located in the N-terminal section; it contains a perfectly conserved hexapeptide. There are two arginine residues in this hexapeptide which could be involved in the binding, via salt bridges, to the carboxyl groups of the propionate side chains of the substrate. This region has been used as a signature pattern. A second signature pattern is based on another well conserved region which is located in the central section of the protein.

Consensus pattern: P-x-W-x-M-R-Q-A-G-R

Consensus pattern: G-F-[STAGCV]-[STAGC]-x-P-[FYW]-T-[LV]-x(2)-Y-x(2)-[AE]-[GK]

[1] Garey J.R., Labbe-Bois R., Chelstowska A., Rytka J., Harrison L., Kushner J., Labbe P.
Eur. J. Biochem. 205:1011-1016(1992).

706. ubiE/COQ5 methyltransferase family signatures

The following methyltransferases have been shown [1] to share regions of similarities: -
Escherichia coli ubiE, which is involved in both ubiquinone and menaquinone biosynthesis
and which catalyzes the S-adenosylmethionine dependent methylation of 2-polyprenyl-6-
methoxy-1,4-benzoquinol into 2-polyprenyl-3- methyl-6-methoxy-1,4-benzoquinol and of
demethylmenaquinol into menaquinol. - Yeast COQ5, a ubiquinone biosynthesis
methlytransferase. - Bacillus subtilis spore germination protein C2 (gene: gercB or gerC2), a
probable menaquinone biosynthesis methlytransferase. - Lactococcus lactis gerC2 homolog. -
Caenorhabditis elegans hypothetical protein ZK652.9. - Leishmania donovani amastigote-
specific protein A41. These are hydrophilic proteins of about 30 Kd (except for ZK652.9
which is 65Kd). They can be picked up in the database by the following patterns.

Consensus pattern: Y-D-x-M-N-x(2)-[LIVM]-S-x(3)-H-x(2)-W

Consensus pattern: R-V-[LIVM]-K-[PV]-G-G-x-[LIVMF]-x(2)-[LIVM]-E-x-S

[1] Lee P.T., Hsu A.Y., Ha H.T., Clarke C.F. J. Bacteriol. 179:1748-1754(1997).

707. Uricase signature

Uricase (urate oxidase) [1] is the peroxisomal enzyme responsible for the degradation of
urate into allantoin. Some species, like primates and birds, have lost the gene for uricase and
are therefore unable to degradeurate. Uricase is a protein of 300 to 400 amino acids. A highly
conserved region located in the central part of the sequence has been used as a signature
pattern.

Consensus pattern: [LV]-x-[LV]-[LIV]-K-[STV]-[ST]-x-[SN]-x-F-x(2)-[FY]-x(4)- [FY]-
x(2)-L-x(5)-R

[1] Motojima K., Kanaya S., Goto S. J. Biol. Chem. 263:16677-16681(1988).

708. Universal stress protein family (Usp)

By a wide range of stress conditions members of the Usp family are predicted to be related to the MADS-box proteins transcript_fact and bind to DNA [2]. Number of members: 39

- 5 [1] Expression and role of the universal stress protein, UspA, of Escherichia coli during growth arrest. Nystrom T, Neidhardt FC; Mol Microbiol 1994; 11:537-544.
[2] Sequence analysis of eukaryotic developmental proteins: ancient and novel domains. Mushegian AR, Koonin EV; Genetics 1996; 144:817-828.

10

709. Ubiquitin domain signature and profile

Ubiquitin [1,2,3] is a protein of seventy six amino acid residues, found in all eukaryotic cells and whose sequence is extremely well conserved from protozoan to vertebrates. It plays a key role in a variety of cellular processes, such as ATP-dependent selective degradation of cellular proteins, maintenance of chromatin structure, regulation of gene expression, stress response and ribosome biogenesis. In most species, there are many genes coding for ubiquitin. However they can be classified into two classes. The first class produces polyubiquitin molecules consisting of exact head to tail repeats of ubiquitin. The number of repeats is variable (up to twelve in a Xenopus gene). In the majority of polyubiquitin precursors, there is a final amino-acid after the last repeat. The second class of genes produces precursor proteins consisting of a single copy of ubiquitin fused to a C-terminal extension protein (CEP). There are two types of CEP proteins and both seem to be ribosomal proteins. Ubiquitin is a globular protein, the last four C-terminal residues (Leu-Arg- Gly-Gly) extending from the compact structure to form a 'tail', important for its function. The latter is mediated by the covalent conjugation of ubiquitin to target proteins, by an isopeptide linkage between the C-terminal glycine and the epsilon amino group of lysine residues in the target proteins. There are a number of proteins which are evolutionary related to ubiquitin: - Ubiquitin-like proteins from baculoviruses as well as in some strains of bovine viral diarrhea viruses (BVDV). These proteins are highly similar to their eukaryotic counterparts. - Mammalian protein GDX [4]. GDX is composed of two domains, a N-terminal ubiquitin-like domain of 74 residues and a C-terminal domain of 83 residues with some similarity with the thyroglobulin hormonogenic site. - Mammalian protein FAU [5]. FAU is a fusion protein which consist of a N-terminal ubiquitin-like protein of 74 residues fused to ribosomal protein

25

30

S30. - Mouse protein NEDD-8 [6], a ubiquitin-like protein of 81 residues. - Human protein BAT3, a large fusion protein of 1132 residues that contains a N-terminal ubiquitin-like domain. - *Caenorhabditis elegans* protein ubl-1 [7]. Ubl-1 is a fusion protein which consist of a N-terminal ubiquitin-like protein of 70 residues fused to ribosomal protein S27A. - Yeast DNA repair protein RAD23 [8]. RAD23 contains a N-terminal domain that seems to be distantly, yet significantly, related to ubiquitin. - Mammalian RAD23-related proteins RAD23A and RAD23B. - Mammalian BCL-2 binding athanogene-1 (BAG-1). BAG-1 is a protein of 274 residues that contains a central ubiquitin-like domain. - Human spliceosome associated protein 114 (SAP 114 or SF3A120). - Yeast protein DSK2, a protein involved in spindle pole body duplication and which contains a N-terminal ubiquitin-like domain. - Human protein CKAP1/TFCB, *Schizosaccharomyces pombe* protein alp11 and *Caenorhabditis elegans* hypothetical protein F53F4.3. These proteins contain a N-terminal ubiquitin domain and a C-terminal CAP-Gly domain. - *Schizosaccharomyces pombe* hypothetical protein SpAC26A3.16. This protein contains a N-terminal ubiquitin domain. - Yeast protein SMT3. - Human ubiquitin-like proteins SMT3A and SMT3B. - Human ubiquitin-like protein SMT3C (also known as PIC1; Ubl1, Sumo-1; Gmp-1 or Sentrin). This protein is involved in targeting ranGAP1 to the nuclear pore complex protein ranBP2. - SMT3-like proteins in plants and *Caenorhabditis elegans*. To identify ubiquitin and related proteins, a pattern has been developed based on conserved positions in the central section of the sequence. A profile was also developed that spans the complete length of the ubiquitin domain.

Consensus pattern: K-x(2)-[LIVM]-x-[DESAK]-x(3)-[LIVM]-[PA]-x(3)-Q-x-[LIVM]-[LIVMC]-[LIVMFY]-x-G-x(4)-[DE]

[1] Jentsch S., Seufert W., Hauser H.-P. *Biochim. Biophys. Acta* 1089:127-139(1991).[2] Monia B.P., Ecker D.J., Croke S.T. *Bio/Technology* 8:209-215(1990).[3] Finley D., Varshavsky A. *Trends Biochem. Sci.* 10:343-347(1985).[4] Filippi M., Tribioli C., Toniolo D. *Genomics* 7:453-457(1990).[5] Olvera J., Wool I.G. *J. Biol. Chem.* 268:17967-17974(1993).[6] Kumar S., Yoshida Y., Noda M. *Biochem. Biophys. Res. Commun.* 195:393-399(1993).[7] Jones D., Candido E.P. *J. Biol. Chem.* 268:19545-19551(1993).[8] Melnick L., Sherman F. *J. Mol. Biol.* 233:372-388(1993).

Domain present in VPS-27, Hrs and STAM. Number of members: 27

711. Vinculin family signatures

Vinculin [1] is a eukaryotic protein that seems to be involved in the attachment of the actin-based microfilaments to the plasma membrane. Vinculin is located at the cytoplasmic side of focal contacts or adhesion plaques. In addition to actin, vinculin interacts with other structural proteins such as talin and alpha-actinins. Vinculin is a large protein of 116 Kd (about a 1000 residues). Structurally the protein consists of an acidic N-terminal domain of about 90 Kd separated from a basic C-terminal domain of about 25 Kd by a proline-rich region of about 50 residues. The central part of the N-terminal domain consists of a variable number (3 in vertebrates, 2 in *Caenorhabditis elegans*) of repeats of a 110 amino acids domain. Catenins [2] are proteins that associate with the cytoplasmic domain of a variety of cadherins. The association of catenins to cadherins produces a complex which is linked to the actin filament network, and which seems to be of primary importance for cadherins cell-adhesion properties. Three different types of catenins seem to exist: alpha, beta, and gamma. Alpha-catenins are proteins of about 100 Kd which are evolutionary related to vinculin. In terms of their structure the most significant differences are the absence, in alpha-catenin, of the repeated domain and of the proline-rich segment. Two signature patterns for this family of proteins have been developed. The first pattern is located in the N-terminal section of both vinculin and alpha-catenins and is part, in vinculin, of a domain that seems to be involved with the interaction with talin. The second pattern is based on a conserved region in the N-terminal part of the repeated domain of vinculin.

Consensus pattern: [KR]-x-[LIVMF]-x(3)-[LIVMA]-x(2)-[LIVM]-x(6)-R-Q-Q-E-L

Consensus pattern: [LIVM]-x-[QA]-A-x(2)-W-[IL]-x-[DN]-P

[1] Otto J.J. Cell Motil. Cytoskeleton 16:1-6(1990). [2] Herrenknecht K., Ozawa M., Eckerskorn C., Lottspeich F., Lenter M., Kemler R. Proc. Natl. Acad. Sci. U.S.A. 88:9156-9160(1991).

712. (Vitellogenin N) Lipoprotein amino terminal region

This family contains regions from: Vitellogenin, Microsomal triglyceride transfer protein and apolipoprotein B-100. These proteins are all involved in lipid transport [1]. This

570

family contains the LV1n chain from lipovitellin, that contains two structural domains.

Number of members: 33

[1] The structural basis of lipid interactions in lipovitellin, a soluble lipoprotein.

Anderson TA, Levitt DG, Banaszak LJ Structure 1998;6:895-909.

5

713. (VMSA) Major surface antigen from hepadnavirus

10 714. ssDNA binding protein (Viral DNA bp)

This protein is found in herpesviruses and is needed for replication.

15 715. (Votage CLC) Voltage gated chloride channels

This family of ion channels contains 10 or 12 transmembrane helices. Each protein forms a single pore. It has been shown that some members of this family form homodimers. These proteins contain two CBS domains.

20 [1] Schmidt-Rose T, Jentsch TJ; J Biol Chem 1997;272:20515-20521.

[2] Zhang J, George AL Jr, Griggs RC, Fouad GT, Roberts J, Kwiecinski H, Connolly AM, Ptacek LJ; Neurology 1996;47:993-998.

25

716. von Willebrand factor type A domain (vwa)

More von Willebrand factor type A domains? Sequence similarities with malaria thrombospondin-related anonymous protein, dihydropyridine-sensitive calcium channel and inter-alpha-trypsin inhibitor.

30

Bork P, Rohde K;

Biochem J 1991;279:908-911.

1. RUGGERI, Z.M. and WARE, J.

von Willebrand factor.

FASEB J. 7 308-316 (1993).

5 2. COLOMBATTI, A., BONALDO, P. and DOLIANA, R.

Type A modules: interacting domains found in several non-fibrillar
collagens and in other extracellular matrix proteins.

MATRIX 13 297-306 (1993).

10 3. PERKINS, S.J., SMITH, K.F., WILLIAMS, S.C., HARIS, P.I., CHAPMAN, D.
and SIM, R.B.

The secondary structure of the von Willebrand factor type A domain in
factor B of human complement by Fourier transform infrared spectroscopy.

Its occurrence in collagen types VI, VII, XII and XIV, the integrins and
other proteins by averaged structure predictions.

J.MOL.BIOL. 238 104-119 (1994).

4. BORK, P. and ROHDE, K.

More von Willebrand factor type A domains? Sequence similarities with
malaria thrombospondin-related anonymous protein, dihydropyridine-
sensitive calcium channel and inter-alpha-trypsin inhibitor.

BIOCHEM.J. 279 908-910 (1991).

5. EDWARDS, Y.J.K. and PERKINS, S.J.

25 The protein fold of the von Willebrand factor type A domain is predicted
to be similar to the open twisted beta-sheet flanked by alpha-helices
found in human ras-p21.

FEBS LETT. 358 283-286 (1995).

30 6. LEE, J.O., RIEU, P., ARNAOUT, M.A. and LIDDINGTON, R.

Crystal structure of the A domain from the alpha subunit of integrin CR3
(CD11b/CD18).

CELL 80 631-638 (1995).

7. QU, A. and LEAHY, D.J.

Crystal structure of the I-domain from the CD11a/CD18 (LFA-1, alpha L beta 2) integrin.

5 PROC.NATL.ACAD.SCI.USA 92 10277-10281 (1995).

The von Willebrand factor is a large multimeric glycoprotein found in blood plasma. Mutant forms are involved in the aetiology of bleeding disorders [1]. In von Willebrand factor, the type A domain (vWF) is the prototype for a protein superfamily. The vWF domain is found in various plasma proteins: complement factors B, C2, CR3 and CR4; the integrins (I-domains); collagen types VI, VII, XII and XIV; and other extracellular proteins [2-4]. Proteins that incorporate vWF domains participate in numerous biological events (e.g., cell adhesion, migration, homing, pattern formation, and signal transduction), involving interaction with a large array of ligands [2].

Secondary structure prediction from 75 aligned vWF sequences has revealed a largely alternating sequence of alpha-helices and beta-strands [3]. Fold recognition algorithms were used to score sequence compatibility with a library of known structures: the vWF domain fold was predicted to be a doubly-wound, open, twisted beta-sheet flanked by alpha-helices [5].

3D structures have been determined for the I-domains of integrins CD11b (with bound magnesium) [6] and CD11a (with bound manganese) [7]. The domain adopts a classic alpha/beta Rossmann fold and contains an unusual metal ion coordination site at its surface. It has been suggested that this site represents a general metal ion-dependent adhesion site (MIDAS) for binding protein ligands [6]. The residues constituting the MIDAS motif in the CD11b and CD11a I-domains are completely conserved, but the manner in which the metal ion is coordinated differs slightly [7].

VWFADOMAIN is a 3-element fingerprint that provides a signature for the vWF domain superfamily. The fingerprint was derived from an initial alignment of 14 sequences. Motif 1 includes the first beta-strand and 3 conserved residues involved in metal ion coordination in I-domains (Asp and 2 serines

in positions 8, 10 and 12, respectively); motif 2 spans strands beta-2 and beta-2'; and motif 3 encodes beta-strand 3 and a conserved Asp (in position 7), which coordinates the metal ion [6,7]. Three iterations on OWL27.0 were required to reach convergence, at which point a true set comprising 56 sequences was identified. Numerous partial matches were also found.

717. (WD40) WD domain, G-beta repeat

The ancient regulatory-protein family of WD-repeat proteins.

Neer EJ, Schmidt CJ, Nambudripad R, Smith TF;
Nature 1994;371:297-300.

Beta-transducin (G-beta) is one of the three subunits (alpha, beta, and gamma) of the guanine nucleotide-binding proteins (G proteins) which act as intermediaries in the transduction of signals generated by transmembrane receptors [1]. The alpha subunit binds to and hydrolyzes GTP; the functions of the beta and gamma subunits are less clear but they seem to be required for the replacement of GDP by GTP as well as for membrane anchoring and receptor recognition.

In higher eukaryotes G-beta exists as a small multigene family of highly conserved proteins of about 340 amino acid residues. Structurally G-beta consists of eight tandem repeats of about 40 residues, each containing a central Trp-Asp motif (this type of repeat is sometimes called a WD-40 repeat). Such a repetitive segment has been shown [E1,2,3,4,5] to exist in a number of other proteins listed below:

- Yeast STE4, a component of the pheromone response pathway. STE4 is a G-beta like protein that associates with GPA1 (G-alpha) and STE18 (G-gamma).
- Yeast MSI1, a negative regulator of RAS-mediated cAMP synthesis. MSI1 is most probably also a G-beta protein.
- Human and chicken protein 12.3. The function of this protein is not known, but on the basis of its similarity to G-beta proteins, it may also function

in signal transduction.

- *Chlamydomonas reinhardtii* gblp. This protein is most probably the homolog of vertebrate protein 12.3.

- Human LIS1, a neuronal protein involved in type-1 lissencephaly [E2].

5 - Mammalian coatamer beta' subunit (beta'-COP), a component of a cytosolic protein complex that reversibly associates with Golgi membranes to form vesicles that mediate biosynthetic protein transport.

10 - Yeast CDC4, essential for initiation of DNA replication and separation of the spindle pole bodies to form the poles of the mitotic spindle.

- Yeast CDC20, a protein required for two microtubule-dependent processes: nuclear movements prior to anaphase and chromosome separation.

- Yeast MAK11, essential for cell growth and for the replication of M1 double-stranded RNA.

15 - Yeast PRP4, a component of the U4/U6 small nuclear ribonucleoprotein with a probable role in mRNA splicing.

- Yeast PWP1, a protein of unknown function.

- Yeast SKI8, a protein essential for controlling the propagation of double-stranded RNA.

20 - Yeast SOF1, a protein required for ribosomal RNA processing which associates with U3 small nucleolar RNA.

- Yeast TUP1 (also known as AER2 or SFL2 or CYC9), a protein which has been implicated in dTMP uptake, catabolite repression, mating sterility, and many other phenotypes.

25 - Yeast YCR57c, an ORF of unknown function from chromosome III.

- Yeast YCR72c, an ORF of unknown function from chromosome III.

- Slime mold coronin, an actin-binding protein.

- Slime mold AAC3, a developmentally regulated protein of unknown function.

30

- Drosophila protein Groucho (formerly known as E(spl); 'enhancer of split'), a protein involved in neurogenesis and that seems to interact with the Notch and Delta proteins.

575

- Drosophila TAF-II-80, a protein that is tightly associated with TFIID.

The number of repeats in the above proteins varies between 5 (PRP4, TUP1, and Groucho) and 8 (G-beta, STE4, MSI1, AAC3, CDC4, PWP1, etc.). In G-beta and G-beta like proteins, the repeats span the entire length of the sequence, while in other proteins, they make up the N-terminal, the central or the C-terminal section.

A signature pattern can be developed from the central core of the domain (positions 9 to 23).

-Consensus pattern: [LIVMSTAC]-[LIVMFYWSTAGC]-[LIMSTAG]-[LIVMSTAGC]-x(2)-[DN]-x(2)-[LIVMWSTAC]-x-[LIVMFSTAG]-W-[DEN]-[LIVMFSTAGCN]

[1] Gilman A.G.

Annu. Rev. Biochem. 56:615-649(1987).

[2] Duronio R.J., Gordon J.I., Boguski M.S.

Proteins 13:41-56(1992).

[3] van der Voorn L., Ploegh H.L.

FEBS Lett. 307:131-134(1992).

[4] Neer E.J., Schmidt C.J., Nambudripad R., Smith T.F.

Nature 371:297-300(1994).

[5] Smith T.F., Gaiatzes C.G., Saxena K., Neer E.J.

Biochemistry In Press(1998).

718. WHEP-TRS domain containing proteins

A conserved domain of 46 amino acids has been shown [1] to exist in a number of higher eukaryote aminoacyl-transfer RNA synthetases. This domain is present one to six times in the following enzymes:

- Mammalian multifunctional aminoacyl-tRNA synthetase. The domain is present

576

three times in a region that separates the N-terminal glutamyl-tRNA synthetase domain from the C-terminal prolyl-tRNA synthetase domain.

- Drosophila multifunctional aminoacyl-tRNA synthetase. The domain is present six times in the intercatalytic region.

5 - Mammalian tryptophanyl-tRNA synthetase. The domain is found at the N-terminal extremity.

- Mammalian, insect, nematode and plant glycyl-tRNA synthetase. The domain is found at the N-terminal extremity [2].

10 - Mammalian histidyl-tRNA synthetase. The domain is found at the N-terminal extremity.

This domain, which is called WHEP-TRS, could contain a central alpha-helical region and may play a role in the association of tRNA-synthetases into multienzyme complexes.

15 A signature pattern based on the first 29 positions of the WHEP-Domain has been developed.

20 -Consensus pattern: [QY]-G-[DNEA]-x-[LIV]-[KR]-x(2)-K-x(2)-[KRNG]-[AS]-x(4)-[LIV]-[DENK]-x(2)-[IV]-x(2)-L-x(3)-K

[1] Cerini C., Kerjan P., Astier M., Gratecos D., Mirande M., Semeriva M.
EMBO J. 10:4267-4277(1991).

25 [2] Nada S., Chang P.K., Dignam J.D.
J. Biol. Chem. 268:7660-7667(1993).

719. (Worm family 8) Putative membrane protein
Analysis of protein domain families in *Caenorhabditis elegans*.

30 Sonnhhammer EL, Durbin R;
Genomics 1997;46:200-216.

This family called family 8 in [1], may be a transmembrane protein
The specific function of this protein is unknown.

720. Xylose isomerase

Xylose isomerase (EC 5.3.1.5) [1] is an enzyme found in microorganisms which catalyzes the interconversion of D-xylose to D-xylulose. It can also isomerize D-ribose to D-ribulose and D-glucose to D-fructose. Xylose isomerase seems to require magnesium for its activity, while cobalt is necessary to stabilize the tetrameric structure of the enzyme. A number of residues are conserved in all known xylose isomerases.

Xylose isomerase also exists in plants [2] where it is homodimeric and is manganese-dependent.

Two signatures patterns for xylose isomerase have been developed. The first one is derived from a stretch of five conserved amino acids that includes a glutamic acid residue known to be one of the four residues involved in the binding of the magnesium ion [3]; this pattern also includes a lysine residue which is involved in the catalytic activity. The second pattern is derived from a conserved region in the N-terminal section of the enzyme that include an histidine residue which has been shown [4] to be involved in the catalytic mechanism of the enzyme.

-Consensus pattern: [LI]-E-P-K-P-x(2)-P

[E is a magnesium ligand]

[K is an active site residue]

-Consensus pattern: [FL]-H-D-x-D-[LIV]-x-[PD]-x-[GDE]

[H is an active site residue]

[1] Dauter Z., Dauter M., Hemker J., Witzel H., Wilson K.S.

FEBS Lett. 247:1-8(1989).

[2] Kristo P.A., Saarelainen R., Fagerstrom R., Aho S., Korhola M.

Eur. J. Biochem. 237:240-246(1996).

[3] Henrick K., Collyer C.A., Blow D.M.

J. Mol. Biol. 208:129-157(1989).

[4] Vangrysperre W., Ampe C., Kersters-Hilderson H., Tempst P.

Biochem. J. 263:195-199(1989).

5

10

15

20

25

30

721. XPG protein signatures. Xeroderma pigmentosum (XP) [1] is a human autosomal recessive disease, characterized by a high incidence of sunlight-induced skin cancer. People's skin cells with this condition are hypersensitive to ultraviolet light, due to defects in the incision step of DNA excision repair. There are a minimum of seven genetic complementation groups involved in this pathway: XP-A to XP-G. The defect in XP-G can be corrected by a 133 Kd nuclear protein called XPG (or XPGC) [2]. XPG belongs to a family of proteins [2,3,4,5,6] that are composed of two main subsets: - Subset 1, to which belongs XPG, RAD2 from budding yeast and rad13 from fission yeast. RAD2 and XPG are single-stranded DNA endonucleases [7,8]. XPG makes the 3' incision in human DNA nucleotide excision repair [9]. - Subset 2, to which belongs mouse and human FEN-1, rad2 from fission yeast, and RAD27 from budding yeast. FEN-1 is a structure-specific endonuclease. In addition to the proteins listed in the above groups, this family also includes: - Fission yeast exo1, a 5'→3' double-stranded DNA exonuclease that could act in a pathway that corrects mismatched base pairs. - Yeast EXO1 (DHS1), a protein with probably the same function as exo1. - Yeast DIN7. Sequence alignment of this family of proteins reveals that similarities are largely confined to two regions. The first is located at the N-terminal extremity (N-region) and corresponds to the first 95 to 105 amino acids. The second region is internal (I-region) and found towards the C-terminus; it spans about 140 residues and contains a highly conserved core of 27 amino acids that includes a conserved pentapeptide (E-A-[DE]-A-[QS]). It is possible that the conserved acidic residues are involved in the catalytic mechanism of DNA excision repair in XPG. The amino acids linking the N- and I-regions are not conserved; indeed, they are largely absent from proteins belonging to the second subset. Two signature patterns have been developed for these proteins. The first corresponds to the central part of the N-region, the second to part of the I-region and includes the putative catalytic core pentapeptide

Consensus pattern: [VI]-[KRE]-P-x-[FYIL]-V-F-D-G-x(2)-[PIL]-x-[LVC]-K-

Consensus pattern: [GS]-[LIVM]-[PER]-[FYS]-[LIVM]-x-A-P-x-E-A-[DE]-[PAS]- [QS]-
[CLM]-

[1] Tanaka K., Wood R.D. Trends Biochem. Sci. 19:83-86(1994).[2] Scherly D., Nospikel
5 T., Corlet J., Ucla C., Bairoch A., Clarkson S.G. Nature 363:182-185(1993).[3] Carr A.M.,
Sheldrick K.S., Murray J.M., Al-Harithy R., Watts F.Z., Lehmann A.R. Nucleic Acids Res.
21:1345-1349(1993).[4] Murray J.M., Tavassoli M., Al-Harithy R., Sheldrick K.S.,
Lehmann A.R., Carr A.M., Watts F.Z. Mol. Cell. Biol. 14:4878-4888(1994).[5] Harrington
10 J.J., Lieber M.R. Genes Dev. 8:1344-1355(1994).[6] Szankasi P., Smith G.R. Science
267:1166-1169(1995).[7] Habraken Y., Sung P., Prakash L., Prakash S. Nature 366:365-
368(1993).[8] O'Donovan A., Scherly D., Clarkson S.G., Wood R.D. J. Biol. Chem.
269:15965-15968(1994).[9] O'Donovan A., Davies A.A., Moggs J.G., West S.C., Wood
R.D. Nature 371:432-435(1994).

722. Xanthine/uracil permeases family

The following transport proteins which are involved in the uptake of xanthine
or uracil are evolutionary related [1]:

- Uric uric acid-xanthine permease (gene uapA) from *Aspergillus nidulans*.
- Purine permease (gene uapC) from *Aspergillus nidulans*.
- Xanthine permease from *Bacillus subtilis* (gene pbuX).
- Uracil permease from *Escherichia coli* (gene uraA) [2] and *Bacillus* (gene
pyrP).
- 25 - Hypothetical protein ycdG from *Escherichia coli*.
- Hypothetical protein ygfO from *Escherichia coli*.
- Hypothetical protein ygfU from *Escherichia coli*.
- Hypothetical protein yicE from *Escherichia coli*.
- Hypothetical protein yunJ from *Bacillus subtilis*.
- 30 - Hypothetical protein yunK from *Bacillus subtilis*.

They are proteins of from 430 to 595 residues that seem to contain 12
transmembrane domains.

The best conserved region which corresponds with what seems to be the tenth transmembrane domain has been selected as a signature pattern.

-Consensus pattern: [LIVM]-P-x-[PASIF]-V-[LIVM]-G-G-x(4)-[LIVM]-[FY]-[GSA]-x-[LIVM]-x(3)-G

[1] Diallinas G., Gorfinkiel L., Arst G., Cecchetto G., Scazzocchio C.
J. Biol. Chem. 270:8610-8622(1995).

[2] Andersen P.S., Frees D., Fast R., Mygind B.
J. Bacteriol. 177:2008-2013(1995).

723. Hypothetical yabO/yceC/sfhB family

The following proteins, which seems to belong to a family of pseudouridine synthases (EC 4.2.1.70) [1] have been shown to share regions of similarities:

- Escherichia coli and Haemophilus influenzae ribosomal large subunit pseudouridine synthase A (gene rluA). It is responsible for synthesis of pseudouridine from uracil-746 IN 23S rRNA.
- Escherichia coli and Haemophilus influenzae ribosomal large subunit pseudouridine synthase C (gene rluC). It is responsible for synthesis of pseudouridine from uracil at positions 955, 2504 and 2580 in 23S rRNA.
- Escherichia coli protein and homologs in other bacteria large subunit pseudouridine synthase D (gene rluD).
- Yeast DRAP deaminase (gene RIB2).
- Escherichia coli hypothetical protein yqcB and HI1435, the corresponding Haemophilus influenzae protein.
- Haemophilus influenzae hypothetical protein HI0042.
- Aquifex aeolicus hypothetical protein AQ_1758.
- Bacillus subtilis hypothetical protein yhcT.
- Bacillus subtilis hypothetical protein yjbO.
- Bacillus subtilis hypothetical protein ylyB.
- Helicobacter pylori hypothetical protein HP0347.
- Helicobacter pylori hypothetical protein HP0745.

581

- Helicobacter pylori hypothetical protein HP0956.
- Mycoplasma genitalium hypothetical protein MG209.
- Mycoplasma genitalium hypothetical protein MG370.
- Synechocystis strain PCC 6803 hypothetical protein slr1592.
- 5 - Synechocystis strain PCC 6803 hypothetical protein slr1629.
- Yeast hypothetical protein YDL036c.
- Yeast hypothetical protein YGR169c.
- Fission yeast hypothetical protein SpAC18B11.02c.
- Caenorhabditis elegans hypothetical protein K07E8.7.

10

These are proteins of from 21 to 50 Kd which contain a number of conserved regions in their central section. They can be picked up in the database by the following highly conserved pattern.

15

-Consensus pattern: [LIVCA]-[NHYT]-R-[LI]-D-x(2)-T-[STA]-G-[LIVAGC]-
[LIVMF](2)-[LIVMFGC]-[SGTACV]

20

[1] Conrad J., Sun D., Englund N., Ofengand J.
J. Biol. Chem. 273:18562-18566(1998).

In addition, the following bacterial proteins, which seems to belong to a family of pseudouridine synthases (EC 4.2.1.70) [1] also have been shown to share regions of similarities:

25

- Escherichia coli and Haemophilus influenzae 16S pseudouridylate 516 synthase (EC 4.2.1.70) (gene: rsuA). This enzyme is responsible for the formation of pseudouridine from uracil-516 in 16S ribosomal RNA.

- Escherichia coli hypothetical protein yciL and HI1199, the corresponding Haemophilus influenzae protein.

30

- Escherichia coli hypothetical protein yjbC.

- Escherichia coli hypothetical protein ymfC and HI0694, the corresponding Haemophilus influenzae protein.

- Aquifex aeolicus hypothetical protein AQ_554.

582

- Aquifex aeolicus hypothetical protein AQ_1464.
- Bacillus subtilis hypothetical protein ypuL.
- Bacillus subtilis hypothetical protein ytzF.
- Borrelia burgdorferi hypothetical protein BB0129.
- 5 - Helicobacter pylori hypothetical protein HP1459.
- Synechocystis strain PCC 6803 hypothetical protein slr0361.
- Synechocystis strain PCC 6803 hypothetical protein slr0612.

These are proteins of from 25 to 40 Kd which contain a number of conserved
10 regions in their central section. They can be picked up in the database by the
following highly conserved pattern.

-Consensus pattern: G-R-L-D-x(2)-[STA]-x-G-[LIVFA]-[LIVMF](3)-[ST]-[DNST]

15 [1] Wrzesinski J., Bakin A., Nurse K., Lane B.G., Ofengand J.
Biochemistry 34:8904-8913(1995).

724. Zinc finger present in dystrophin, CBP/p300

20 ZZ in dystrophin binds calmodulin

Putative zinc finger; binding not yet shown.

725. Zinc carboxypeptidase

25 There are a number of different types of zinc-dependent carboxypeptidases (EC
3.4.17.-) [1,2]. All these enzymes seem to be structurally and functionally
related. The enzymes that belong to this family are listed below.

- 30 - Carboxypeptidase A1 (EC 3.4.17.1), a pancreatic digestive enzyme that can
removes all C-terminal amino acids with the exception of Arg, Lys and Pro.
- Carboxypeptidase A2 (EC 3.4.17.15), a pancreatic digestive enzyme with a
specificity similar to that of carboxypeptidase A1, but with a preference
for bulkier C-terminal residues.

583

- Carboxypeptidase B (EC 3.4.17.2), also a pancreatic digestive enzyme, but that preferentially removes C-terminal Arg and Lys.
- Carboxypeptidase N (EC 3.4.17.3) (also known as arginine carboxypeptidase), a plasma enzyme which protects the body from potent vasoactive and inflammatory peptides containing C-terminal Arg or Lys (such as kinins or anaphylatoxins) which are released into the circulation.
- Carboxypeptidase H (EC 3.4.17.10) (also known as enkephalin convertase or carboxypeptidase E), an enzyme located in secretory granules of pancreatic islets, adrenal gland, pituitary and brain. This enzyme removes residual C-terminal Arg or Lys remaining after initial endoprotease cleavage during prohormone processing.
- Carboxypeptidase M (EC 3.4.17.12), a membrane bound Arg and Lys specific enzyme.
It is ideally situated to act on peptide hormones at local tissue sites where it could control their activity before or after interaction with specific plasma membrane receptors.
- Mast cell carboxypeptidase (EC 3.4.17.1), an enzyme with a specificity to carboxypeptidase A, but found in the secretory granules of mast cells.
- Streptomyces griseus carboxypeptidase (Cpase SG) (EC 3.4.17.-) [3], which combines the specificities of mammalian carboxypeptidases A and B.
- Thermoactinomyces vulgaris carboxypeptidase T (EC 3.4.17.18) (CPT) [4], which also combines the specificities of carboxypeptidases A and B.
- AEBP1 [5], a transcriptional repressor active in preadipocytes. AEBP1 seems to regulate transcription by cleavage of other transcriptional proteins.
- Yeast hypothetical protein YHR132c.

All of these enzymes bind an atom of zinc. Three conserved residues are implicated in the binding of the zinc atom: two histidines and a glutamic acid. Two signature patterns which contain these three zinc-ligands have been derived.

-Consensus pattern: [PK]-x-[LIVMFY]-x-[LIVMFY]-x(4)-H-[STAG]-x-E-x-[LIVM]-[STAG]-x(6)-[LIVMFYTA]
[H and E are zinc ligands]

-Consensus pattern: H-[STAG]-x(3)-[LIVME]-x(2)-[LIVMFYW]-P-[FYW]
[H is a zinc ligand]

[1] Tan F., Chan S.J., Steiner D.F., Schilling J.W., Skidgel R.A.

J. Biol. Chem. 264:13165-13170(1989).

[2] Reynolds D.S., Stevens R.L., Gurley D.S., Lane W.S., Austen K.F.,
Serafin W.E.

J. Biol. Chem. 264:20094-20099(1989).

[3] Narahashi Y.

J. Biochem. 107:879-886(1990).

[4] Teplyakov A., Polyakov K., Obmolova G., Strokopytov B., Kuranova I.,
Osterman A.L., Grishin N.V., Smulevitch S.V., Zagnitko O.P.,
Galperina O.V., Matz M.V., Stepanov V.M.
Eur. J. Biochem. 208:281-288(1992).

[5] He G.-P., Muise A., Li A.W., Ro H.-S.
Nature 378:92-96(1995).

[6] Hourdou M.-L., Guinand M., Vacheron M.J., Michel G., Denoroy L.,
Duez C.M., Englebert S., Joris B., Weber G., Ghuysen J.-M.
Biochem. J. 292:563-570(1993).

[7] Rawlings N.D., Barrett A.J.
Meth. Enzymol. 248:183-228(1995).

726. Zinc finger, C2H2 type

The C2H2 zinc finger is the classical zinc finger domain.

The two conserved cysteines and histidines co-ordinate a
zinc ion. The following pattern describes the zinc finger.

#-X-C-X(1-5)-C-X3-#-X5-#-X2-H-X(3-6)-[H/C]

Where X can be any amino acid, and numbers in brackets

indicate the number of residues. The positions marked # are
those that are important for the stable fold of the zinc
finger. The final position can be either his or cys.

The C2H2 zinc finger is composed of two short beta strands

585

followed by an alpha helix. The amino terminal part of the helix binds the major groove in DNA binding zinc fingers.

'Zinc finger' domains [1-5] are nucleic acid-binding protein structures first identified in the *Xenopus* transcription factor TFIIIA. These domains have since been found in numerous nucleic acid-binding proteins. A zinc finger domain is composed of 25 to 30 amino-acid residues. There are two cysteine or histidine residues at both extremities of the domain, which are involved in the tetrahedral coordination of a zinc atom. It has been proposed that such a domain interacts with about five nucleotides. A schematic representation of a zinc finger domain is shown below:

```

      x x
      x  x
      x x
      x x
      x x
      x x
      C  H
      x \ / x
      x  Zn  x
      x /  \ x
      C  H
      x x x x x x x x x x

```

Many classes of zinc fingers are characterized according to the number and positions of the histidine and cysteine residues involved in the zinc atom coordination. In the first class to be characterized, called C2H2, the first pair of zinc coordinating residues are cysteines, while the second pair are histidines. A number of experimental reports have demonstrated the zinc-dependent DNA or RNA binding property of some members of this class.

Some of the proteins known to include C2H2-type zinc fingers are listed below.

The number of zinc finger regions found in each of these proteins are indicated between brackets; a '+' symbol indicates that only partial sequence data is available and that additional finger domains may be present.

- 5 - *Saccharomyces cerevisiae*: ACE2 (3), ADR1 (2), AZF1 (4), FZF1 (5), MIG1 (2), MSN2 (2), MSN4 (2), RGM1 (2), RIM1 (3), RME1 (3), SFP1 (2), SSL1 (1), STP1 (3), SWI5 (3), VAC1 (1) and ZMS1 (2).
- *Emmericella nidulans*: brlA (2), creA (2).
- *Drosophila*: AEF-1 (4), Cf2 (7), ci-D (5), Disconnected (2), Escargot (5),
10 Glass (5), Hunchback (6), Kruppel (5), Kruppel-H (4+), Odd-skipped (4), Odd-paired (4), Pep (3), Snail (5), Spalt-major (7), Serependity locus beta (6), delta (7), h-1 (8), Suppressor of hairy wing su(Hw) (12), Suppressor of variegation suvar(3)7 (5), Teashirt (3) and Tramtrack (2).
- *Xenopus*: transcription factor TFIIIA (9), p43 from RNP particle (9), Xfin (37 !!), Xsna (5), gastrula XlcGF5.1 to XlcGF71.1 (from 4+ to 11+), Oocyte XlcOF2 to XlcOF22 (from 7 to 12).
- Mammalian: basonuclin (6), BCL-6/LAZ-3 (6), erythroid krueppel-like transcription factor (3), transcription factors Sp1 (3), Sp2 (3), Sp3 (3) and Sp(4) 3, transcriptional repressor YY1 (4), Wilms' tumor protein (4),
15 EGR1/Krox24 (3), EGR2/Krox20 (3), EGR3/Pilot (3), EGR4/AT133 (4), Evi-1 (10), GLI1 (5), GLI2 (4+), GLI3 (3+), HIV-EP1/ZNF40 (4), HIV-EP2 (2), KR1 (9+), KR2 (9), KR3 (15+), KR4 (14+), KR5 (11+), HF.12 (6+), REX-1 (4), ZfX (13), ZfY (13), Zfp-35 (18), ZNF7 (15), ZNF8 (7), ZNF35 (10), ZNF42/MZF-1 (13), ZNF43 (22), ZNF46/Kup (2), ZNF76 (7), ZNF91 (36), ZNF133 (3).

25

In addition to the conserved zinc ligand residues it has been shown [6] that a number of other positions are also important for the structural integrity of the C2H2 zinc fingers. The best conserved position is found four residues after the second cysteine; it is generally an aromatic or aliphatic residue.

30

-Consensus pattern: C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H
[The two C's and two H's are zinc ligands]

- [1] Klug A., Rhodes D.
Trends Biochem. Sci. 12:464-469(1987).
- [2] Evans R.M., Hollenberg S.M.
Cell 52:1-3(1988).
- 5 [3] Payre F., Vincent A.
FEBS Lett. 234:245-250(1988).
- [4] Miller J., McLachlan A.D., Klug A.
EMBO J. 4:1609-1614(1985).
- [5] Berg J.M.
10 Proc. Natl. Acad. Sci. U.S.A. 85:99-102(1988).
- [6] Rosenfeld R., Margalit H.
J. Biomol. Struct. Dyn. 11:557-570(1993).

15 727. Zinc finger, C3HC4 type (RING finger)

A number of eukaryotic and viral proteins contain a conserved cysteine-rich domain of 40 to 60 residues (called C3HC4 zinc-finger or 'RING' finger) [1] that binds two atoms of zinc, and is probably involved in mediating protein-protein interactions. The 3D structure of the zinc ligation system is unique to the RING domain and is referred to as the "cross-brace" motif. The spacing of the cysteines in such a domain is C-x(2)-C-x(9 to 39)-C-x(1 to 3)-H-x(2 to 3)-C-x(2)-C-x(4 to 48)-C-x(2)-C.

Proteins currently known to include the C3HC4 domain are listed below
25 (references are only provided for recently determined sequences).

- Mammalian V(D)J recombination activating protein (gene RAG1). RAG1 activates the rearrangement of immunoglobulin and T-cell receptor genes.
- Mouse rpt-1. Rpt-1 is a trans-acting factor that regulates gene expression
30 directed by the promoter region of the interleukin-2 receptor alpha chain or the LTR promoter region of HIV-1.
- Human rfp. Rfp is a developmentally regulated protein that may function in male germ cell development. Recombination of the N-terminal section of rfp

with a protein tyrosine kinase produces the ret transforming protein.

- Human 52 Kd Ro/SS-A protein. A protein of unknown function from the Ro/SS-A ribonucleoprotein complex. Sera from patients with systemic lupus erythematosus or primary Sjogren's syndrome often contain antibodies that react with the Ro proteins.
- Human histocompatibility locus protein RING1.
- Human PML, a probable transcription factor. Chromosomal translocation of PML with retinoic receptor alpha creates a fusion protein which is the cause of acute promyelocytic leukemia (APL).
- Mammalian breast cancer type 1 susceptibility protein (BRCA1) [E1].
- Mammalian cbl proto-oncogene.
- Mammalian bmi-1 proto-oncogene.
- Vertebrate CDK-activating kinase (CAK) assembly factor MAT1, a protein that stabilizes the complex between the CDK7 kinase and cyclin H (MAT1 stands for 'Menage A Trois').
- Mammalian mel-18 protein. Mel-18 which is expressed in a variety of tumor cells is a transcriptional repressor that recognizes and bind a specific DNA sequence.
- Mammalian peroxisome assembly factor-1 (PAF-1) (PMP35), which is somewhat involved in the biogenesis of peroxisomes. In humans, defects in PAF-1 are responsible for a form of Zellweger syndrome, an autosomal recessive disorder associated with peroxisomal deficiencies.
- Human MAT1 protein, which interacts with the CDK7-cyclin H complex.
- Human RING1 protein.
- Xenopus XNF7 protein, a probable transcription factor.
- Trypanosoma protein ESAG-8 (T-LR), which may be involved in the postranscriptional regulation of genes in VSG expression sites or may interact with adenylate cyclase to regulate its activity.
- Drosophila proteins Posterior Sex Combs (Psc) and Suppressor two of zeste (Su(z)2). The two proteins belong to the Polycomb group of genes needed to maintain the segment-specific repression of homeotic selector genes.
- Drosophila protein male-specific msl-2, a DNA-binding protein which is involved in X chromosome dosage compensation (the elevation of

589

transcription of the male single X chromosome).

- Arabidopsis thaliana protein COP1 which is involved in the regulation of photomorphogenesis.

- Fungal DNA repair proteins RAD5, RAD16, RAD18 and rad8.

5 - Herpesviruses trans-acting transcriptional protein ICP0/IE110. This protein which has been characterized in many different herpesviruses is a trans-activator and/or -repressor of the expression of many viral and cellular promoters.

- Baculoviruses protein CG30.

10 - Baculoviruses major immediate early protein (PE-38).

- Baculoviruses immediate-early regulatory protein IE-N/IE-2.

- Caenorhabditis elegans hypothetical proteins F54G8.4, R05D3.4 and T02C1.1.

- Yeast hypothetical proteins YER116c and YKR017c.

5 The central region of the domain was selected as a signature pattern for the C3HC4 finger.

-Consensus pattern: C-x-H-x-[LIVMFY]-C-x(2)-C-[LIVMYA]

20 [1] Borden K.L.B., Freemont P.S.

Curr. Opin. Struct. Biol. 6:395-401(1996).

728. Zinc finger C-x8-C-x5-C-x3-H type (and similar).

25

729. Zinc finger, CCHC class

A family of CCHC zinc fingers, mostly from retroviral gag proteins (nucleocapsid). Prototype structure is from HIV.

30 Also contains members involved in eukaryotic gene regulation, such as C. elegans GLH-1.

Structure is an 18-residue zinc finger; no examples of indels in the alignment.

730. Zn-finger in Ran binding protein and others.

5

731. AN1-like Zinc finger

Zinc finger at the C-terminus of An1 Swiss:Q91889, a ubiquitin-like protein in *Xenopus laevis*. The following pattern describes the zinc finger. C-X2-C-X(9-12)-C-X(1-2)-C-X4-C-
10 X2-H-X5-H-X-C Where X can be any amino acid, and numbers in brackets indicate the number of residues.

[1] Linnen JM, Bailey CP, Weeks DL; Gene 1993;128:181-188.

5

732. 14-3-3 proteins

Structure of a 14-3-3 protein and implications for coordination of multiple signalling pathways.

Xiao B, Smerdon SJ, Jones DH, Dodson GG, Soneji Y, Aitken A, Gamblin SJ;
20 Nature 1995;376:188-191.

Crystal structure of the zeta isoform of the 14-3-3 protein.

Liu D, Bienkowska J, Petosa C, Collier RJ, Fu H, Liddington R;
Nature 1995;376:191-194.

25 Interaction of 14-3-3 with signaling proteins is mediated by the recognition of phosphoserine.

Muslin AJ, Tanner JW, Allen PM, Shaw AS;
Cell 1996;84:889-897.

30 The 14-3-3 protein binds its target proteins with a common site located towards the C-terminus.

Ichimura T, Ito M, Itagaki C, Takahashi M, Horigome T, Omata S, Ohno S, Isobe T

FEBS Lett 1997;413:273-276.

Molecular evolution of the 14-3-3 protein family.

Wang W, Shakes DC

5 J Mol Evol 1996;43:384-398.

Function of 14-3-3 proteins.

Jin DY, Lyu MS, Kozak CA, Jeang KT

Nature 1996;382:308-308.

10 The 14-3-3 proteins [1,2,3] are a family of closely related acidic homodimeric proteins of about 30 Kd which were first identified as being very abundant in mammalian brain tissues and located preferentially in neurons. The 14-3-3 proteins seem to have multiple biological activities and play a key role in signal transduction pathways and the cell cycle. They interact with kinases
15 such as PKC or Raf-1; they seem to also function as protein-kinase dependent activators of tyrosine and tryptophan hydroxylases and in plants they are associated with a complex that binds to the G-box promoter elements.

20 The 14-3-3 family of proteins are ubiquitously found in all eukaryotic species studied and have been sequenced in fungi (yeast BMH1 and BMH2, fission yeast rad24 and rad25), plants, Drosophila, and vertebrates. The sequences of the 14-3-3 proteins are extremely well conserved. Two highly conserved regions have been selected as signature patterns: the first is a peptide of 11 residues
25 located in the N-terminal section; the second, a 20 amino acid region located in the C-terminal section.

-Consensus pattern: R-N-L-[LIV]-S-[VG]-[GA]-Y-[KN]-N-[IVA]

-Consensus pattern: Y-K-[DE]-S-T-L-I-[IM]-Q-L-[LF]-[RHC]-D-N-[LF]-T-[LS]-W-
[TAN]-[SAD]

30

[1] Aitken A.

Trends Biochem. Sci. 20:95-97(1995).

[2] Morrison D.

Science 266:56-57(1994).

[3] Xiao B., Smerdon S.J., Jones D.H., Dodson G.G., Soneji Y., Aitken A.,
Gamblin S.J.

Nature 376:188-191(1995).

5

733. D-isomer specific 2-hydroxyacid dehydrogenases (2 Hacid DH)

This Pfam covers the Formate dehydrogenase, D-glycerate dehydrogenase and D-lactate dehydrogenase families in SCOP. A number of NAD-dependent 2-hydroxyacid dehydrogenases which seem to be specific for the D-isomer of their substrate have been shown [1,2,3,4] to be functionally and structurally related. These enzymes are listed below.

- D-lactate dehydrogenase (EC 1.1.1.28), a bacterial enzyme which catalyzes the reduction of D-lactate to pyruvate.
- D-glycerate dehydrogenase (EC 1.1.1.29) (NADH-dependent hydroxypyruvate reductase), a plant leaf peroxisomal enzyme that catalyzes the reduction of hydroxypyruvate to glycerate. This reaction is part of the glycolate pathway of photorespiration.
- D-glycerate dehydrogenase from the bacteria *Hyphomicrobium methylovorum* and *Methylobacterium extorquens*.
- 3-phosphoglycerate dehydrogenase (EC 1.1.1.95), a bacterial enzyme that catalyzes the oxidation of D-3-phosphoglycerate to 3-phosphohydroxypyruvate. This reaction is the first committed step in the 'phosphorylated' pathway of serine biosynthesis.
- Erythronate-4-phosphate dehydrogenase (EC 1.1.1.-) (gene *pdxB*), a bacterial enzyme involved in the biosynthesis of pyridoxine (vitamin B6).
- D-2-hydroxyisocaproate dehydrogenase (EC 1.1.1.-) (D-hicDH), a bacterial enzyme that catalyzes the reversible and stereospecific interconversion between 2-ketocarboxylic acids and D-2-hydroxy-carboxylic acids.
- Formate dehydrogenase (EC 1.2.1.2) (FDH) from the bacteria *Pseudomonas* sp. 101 and various fungi [5].
- Vancomycin resistance protein *vanH* from *Enterococcus faecium*; this protein is a D-specific alpha-keto acid dehydrogenase involved in the formation of a peptidoglycan which does not terminate by D-alanine thus preventing vancomycin binding.
- *Escherichia coli* hypothetical protein *ycdW*.

10

15

20

25

30

35

593

- Escherichia coli hypothetical protein yiaE.
- Haemophilus influenzae hypothetical protein HI1556.
- Yeast hypothetical protein YER081w.
- Yeast hypothetical protein YIL074w.

5 All these enzymes have similar enzymatic activities and are structurally related. Three of the most conserved regions of these proteins have been selected to develop patterns. The first pattern is based on a glycine-rich region located in the central section of these enzymes; this region probably corresponds to the NAD-binding domain. The two other patterns contain a number of conserved charged residues, some of which may play a role in the catalytic

10 mechanism.

-Consensus pattern: [LIVMA]-[AG]-[IVT]-[LIVMFY]-[AG]-x-G-[NHKROGSAC]-[LIV]-G-x(13,14)-[LIVfMT]-x(2)-[FYwCTH]-[DNSTK]

-Consensus pattern: [LIVMFYWA]-[LIVFYWC]-x(2)-[SAC]-[DNQHR]-[IVFA]-[LIVF]-x-[LIVF]-[HNI]-x-P-x(4)-[STN]-x(2)-[LIVMF]-x-[GSDN]

-Consensus pattern: [LMFATC]-[KPO]-x-[GSTDN]-x-[LIVMFYWR]-[LIVMFYW](2)-N-x-[STAGC]-R-[GP]-x-[LIVH]-[LIVMC]-[DNV]

[1] Grant G.A. Biochem. Biophys. Res. Commun. 165:1371-1374(1989).

[2] Kochhar S., Hunziker P., Leong-Morgenthaler P.M., Hottinger H. Biochem. Biophys. Res. Commun. 184:60-66(1992).

[3] Ohta T., Taguchi H. J. Biol. Chem. 266:12588-12594(1991).

[4] Goldberg J.D., Yoshida T., Brick P. J. Mol. Biol. 236:1123-1140(1994).

[5] Popov V.O., Lamzin V.S. Biochem. J. 301:625-643(1994).

734. 2-oxo acid dehydrogenases acyltransferase (catalytic domain)

Refined crystal structure of the catalytic domain of dihydrolipoyl transacetylase (E2P) from azotobacter vineelandii at 2.6 angstroms

30 resolution.

Mattevi A, Obmolova G, Kalk KH, Westphal AH, De Kok A, Hol WG;
J Mol Biol 1993;230:1183-1199.

These proteins contain one to three copies of a lipoyl binding domain

followed by the catalytic domain.

735. 3-beta hydroxysteroid dehydrogenase/isomerase family

Structure and tissue-specific expression of 3
beta-hydroxysteroid dehydrogenase/5-ene-4-ene isomerase
genes in human and rat classical and peripheral
steroidogenic tissues.

Labrie F, Simard J, Luu-The V, Pelletier G, Belanger A,
Lachance Y, Zhao HF, Labrie C, Breton N, de Launoit Y, et al
J Steroid Biochem Mol Biol 1992;41:421-435.

The enzyme 3 beta-hydroxysteroid dehydrogenase/5-ene-4-ene
isomerase (3 beta-HSD) catalyzes the oxidation and isomerization
of 5-ene-3 beta-hydroxypregnene and 5-ene-hydroxyandrostene
steroid precursors into the corresponding 4-ene-ketosteroids necessary
for the formation of all classes of steroid hormones.

736. 3-hydroxyacyl-CoA dehydrogenase

This family also includes lambda crystallin.
Structure of L-3-hydroxyacyl-coenzyme A dehydrogenase:
preliminary chain tracing at 2.8-A resolution.

Birktoft JJ, Holden HM, Hamlin R, Xuong NH, Banaszak LJ;
Proc Natl Acad Sci U S A 1987;84:8262-8266.

3-hydroxyacyl-CoA dehydrogenase (EC 1.1.1.35) (HCDH) [1] is an enzyme involved
in fatty acid metabolism, it catalyzes the reduction of 3-hydroxyacyl-CoA to
3-oxoacyl-CoA. Most eukaryotic cells have 2 fatty-acid beta-oxidation systems,
one located in mitochondria and the other in peroxisomes. In peroxisomes
3-hydroxyacyl-CoA dehydrogenase forms, with enoyl-CoA hydratase (ECH) and
3,2-trans-enoyl-CoA isomerase (ECI) a multifunctional enzyme where the N-
terminal domain bears the hydratase/isomerase activities and the C-terminal
domain the dehydrogenase activity. There are two mitochondrial enzymes: one

which is monofunctional and the other which is, like its peroxisomal counterpart, multifunctional.

In *Escherichia coli* (gene *fadB*) and *Pseudomonas fragi* (gene *faoA*) HCDH is part of a multifunctional enzyme which also contains an ECH/ECI domain as well as a 3-hydroxybutyryl-CoA epimerase domain [2].

The other proteins structurally related to HCDH are:

- Bacterial 3-hydroxybutyryl-CoA dehydrogenase (EC 1.1.1.157) which reduces 3-hydroxybutanoyl-CoA to acetoacetyl-CoA [3].
- Eye lens protein lambda-crystallin [4], which is specific to lagomorphes (such as rabbit).

There are two major region of similarities in the sequences of proteins of the HCDH family, the first one located in the N-terminal, corresponds to the NAD-binding site, the second one is located in the center of the sequence. A signature pattern has been derived from this central region.

-Consensus pattern: [DNE]-x(2)-[GA]-F-[LIVMFY]-x-[NT]-R-x(3)-[PA]-[LIVMFY](2)-x(5)-[LIVMFYCT]-[LIVMFY]-x(2)-[GV]

[1] Birktoff J.J., Holden H.M., Hamlin R., Xuong N.-H., Banaszak L.J.
Proc. Natl. Acad. Sci. U.S.A. 84:8262-8266(1987).

[2] Nakahigashi K., Inokuchi H.
Nucleic Acids Res. 18:4937-4937(1990).

[3] Mullany P., Clayton C.L., Pallen M.J., Slone R., Al-Saleh A.,
Tabaqchali S.
FEMS Microbiol. Lett. 124:61-67(1994).

[4] Mulders J.W.M., Hendriks W., Blankesteyn W.M., Bloemendal H.,
de Jong W.W.
J. Biol. Chem. 263:15462-15466(1988).

737. 60s Acidic ribosomal protein

Proteins P1, P2, and P0, components of the eukaryotic ribosome stalk. New structural and functional aspects.

- 5 Remacha M, Jimenez-Diaz A, Santos C, Briones E, Zambrano R, Rodriguez Gabriel MA, Guarinos E, Ballesta JP; Biochem Cell Biol 1995;73:959-968.
This family includes archaeobacterial L12, eukaryotic P0, P1 and P2.

10

738. 6-phosphogluconate dehydrogenases

6-phosphogluconate dehydrogenase (EC 1.1.1.44) (6PGD) catalyzes the third step in the hexose monophosphate shunt, the decarboxylating reduction of 6-phosphogluconate in to ribulose 5-phosphate.

5

Prokaryotic and eukaryotic 6PGD are proteins of about 470 amino acids whose sequence are highly conserved [1]. A region which has been shown [2], from studies of the sheep 6PGD tertiary structure, to be involved in the binding of 6-phosphogluconate has been selected as a signature pattern.

20

-Consensus pattern: [LIVM]-x-D-x(2)-[GA]-[NQS]-K-G-T-G-x-W

[1] Reizer A., Deutscher J., Saier M.H. Jr., Reizer J.

Mol. Microbiol. 5:1081-1089(1991).

25 [2] Adams M.J., Archibald I.G., Bugg C.E., Carne A., Gover S.,

Helliwell J.R., Pickersgill R.W., White S.W.

EMBO J. 2:1009-1014(1983).

- 30 739. (7tm 1) G-protein coupled receptors [1 to 4,E1,E2] (also called R7G) are an extensive group of hormones, neurotransmitters, odorants and light receptors which transduce extracellular signals by interaction with guanine nucleotide-binding (G) proteins. The receptors that are currently known to belong to this

family are listed below.

- 5-hydroxytryptamine (serotonin) 1A to 1F, 2A to 2C, 4, 5A, 5B, 6 and 7 [5].
- Acetylcholine, muscarinic-type, M1 to M5.
- 5 - Adenosine A1, A2A, A2B and A3 [6].
- Adrenergic alpha-1A to -1C; alpha-2A to -2D; beta-1 to -3 [7].
- Angiotensin II types I and II.
- Bombesin subtypes 3 and 4.
- Bradykinin B1 and B2.
- 10 - c3a and C5a anaphylatoxin.
- Cannabinoid CB1 and CB2.
- Chemokines C-C CC-CKR-1 to CC-CKR-8.
- Chemokines C-X-C CXC-CKR-1 to CXC-CKR-4.
- Cholecystokinin-A and cholecystokinin-B/gastrin.
- 15 - Dopamine D1 to D5 [8].
- Endothelin ET-a and ET-b [9].
- fMet-Leu-Phe (fMLP) (N-formyl peptide).
- Follicle stimulating hormone (FSH-R) [10].
- Galanin.
- 20 - Gastrin-releasing peptide (GRP-R).
- Gonadotropin-releasing hormone (GNRH-R).
- Histamine H1 and H2 (gastric receptor I).
- Lutropin-choriogonadotropic hormone (LSH-R) [10].
- Melanocortin MC1R to MC5R.
- 25 - Melatonin.
- Neuromedin B (NMB-R).
- Neuromedin K (NK-3R).
- Neuropeptide Y types 1 to 6.
- Neurotensin (NT-R).
- 30 - Octopamine (tyramine), from insects.
- Odorants [11].
- Opioids delta-, kappa- and mu-types [12].
- Oxytocin (OT-R).

- Platelet activating factor (PAF-R).
- Prostacyclin.
- Prostaglandin D2.
- Prostaglandin E2, EP1 to EP4 subtypes.
- 5 - Prostaglandin F2.
- Purinoreceptors (ATP) [13].
- Somatostatin types 1 to 5.
- Substance-K (NK-2R).
- Substance-P (NK-1R).
- 10 - Thrombin.
- Thromboxane A2.
- Thyrotropin (TSH-R) [10].
- Thyrotropin releasing factor (TRH-R).
- Vasopressin V1a, V1b and V2.
- 15 - Visual pigments (opsins and rhodopsin) [14].
- Proto-oncogene mas.
- A number of orphan receptors (whose ligand is not known) from mammals and birds.
- *Caenorhabditis elegans* putative receptors C06G4.5, C38C10.1, C43C3.2, T27D1.3 and ZC84.4.
- 20 - Three putative receptors encoded in the genome of cytomegalovirus: US27, US28, and UL33.
- ECRF3, a putative receptor encoded in the genome of herpesvirus saimiri.

25 The structure of all these receptors is thought to be identical. They have seven hydrophobic regions, each of which most probably spans the membrane. The N-terminus is located on the extracellular side of the membrane and is often glycosylated, while the C-terminus is cytoplasmic and generally phosphorylated. Three extracellular loops alternate with three intracellular

30 loops to link the seven transmembrane regions. Most, but not all of these receptors, lack a signal peptide. The most conserved parts of these proteins are the transmembrane regions and the first two cytoplasmic loops. A conserved acidic-Arg-aromatic triplet is present in the N-terminal extremity of the

second cytoplasmic loop [15] and could be implicated in the interaction with G proteins.

To detect this widespread family of proteins, a pattern that contains the conserved triplet and that also spans the major part of the third transmembrane helix has been developed.

-Consensus pattern: [GSTALIVMFYWC]-[GSTANCPDE]-{EDPKRH}-x(2)-
[LIVMNQGA]-x(2)-
[LIVMFT]-[GSTANC]-[LIVMFYWSTAC]-[DENH]-R-[FYWCSH]-x(2)-
[LIVM]

[1] Strosberg A.D.

Eur. J. Biochem. 196:1-10(1991).

[2] Kerlavage A.R.

Curr. Opin. Struct. Biol. 1:394-401(1991).

[3] Probst W.C., Snyder L.A., Schuster D.I., Brosius J., Sealfon S.C.

DNA Cell Biol. 11:1-20(1992).

[4] Savarese T.M., Fraser C.M.

Biochem. J. 283:1-9(1992).

[5] Branchek T.

Curr. Biol. 3:315-317(1993).

[6] Stiles G.L.

J. Biol. Chem. 267:6451-6454(1992).

[7] Friell T., Kobilka B.K., Lefkowitz R.J., Caron M.G.

Trends Neurosci. 11:321-324(1988).

[8] Stevens C.F.

Curr. Biol. 1:20-22(1991).

[9] Sakurai T., Yanagisawa M., Masaki T.

Trends Pharmacol. Sci. 13:103-107(1992).

[10] Salesse R., Remy J.J., Levin J.M., Jallal B., Garnier J.

Biochimie 73:109-120(1991).

[11] Lancet D., Ben-Arie N.

Curr. Biol. 3:668-674(1993).

[12] Uhl G.R., Childers S., Pasternak G.

Trends Neurosci. 17:89-93(1994).

[13] Barnard E.A., Burnstock G., Webb T.E.

5 Trends Pharmacol. Sci. 15:67-70(1994).

[14] Applebury M.L., Hargrave P.A.

Vision Res. 26:1881-1895(1986).

[15] Attwood T.K., Eliopoulos E.E., Findlay J.B.C.

Gene 98:153-159(1991).

10

(7tm 1) Visual pigments (opsins) retinal binding site

Visual pigments [1,2] are the light-absorbing molecules that mediate vision.

They consist of an apoprotein, opsin, covalently linked to the chromophore

cis-retinal. Vision is effected through the absorption of a photon by cis-

15 retinal which is isomerized to trans-retinal. This isomerization leads to a

change of conformation of the protein. Opsins are integral membrane proteins

with seven transmembrane regions that belong to family 1 of G-protein coupled
receptors.

20 In vertebrates four different pigments are generally found. Rod cells, which
mediate vision in dim light, contain the pigment rhodopsin. Cone cells, which
function in bright light, are responsible for color vision and contain three
or more color pigments (for example, in mammals: red, blue and green).

25 In Drosophila, the eye is composed of 800 facets or ommatidia. Each
ommatidium contains eight photoreceptor cells (R1-R8): the R1 to R6 cells are
outer cells, R7 and R8 inner cells. Each of the three types of cells (R1-R6,
R7 and R8) expresses a specific opsin.

30 Proteins evolutionary related to opsins include squid retinochrome, also known
as retinal photoisomerase, which converts various isomers of retinal into 11-
cis retinal and mammalian retinal pigment epithelium (RPE) RGR [3], a protein
that may also act in retinal isomerization.

The attachment site for retinal in the above proteins is a conserved lysine residue in the middle of the seventh transmembrane helix. The pattern that had been developed includes this residue.

5

-Consensus pattern: [LIVMWAC]-[PGC]-x(3)-[SAC]-K-[STALIMR]-[GSACPNV]-
[STACP]-
x(2)-[DENF]-[AP]-x(2)-[IY]
[K is the retinal binding site]

10

- [1] Applebury M.L., Hargrave P.A.
Vision Res. 26:1881-1895(1986).
- [2] Fryxell K.J., Meyerowitz E.M.
J. Mol. Evol. 33:367-378(1991).
- [3] Shen D., Jiang M., Hao W., Tao L., Salazar M., Fong H.K.W.
Biochemistry 33:13117-13125(1994).

15

The following descriptions of protein family functions are not provided by the Pfam or Prosite databases.

20

740. BAH

BAH domain. Number of members: 65

25

[1] Medline: 97074677. Molecular cloning of polybromo, a nuclear protein containing multiple domains including five bromodomains, a truncated HMG-box, and two repeats of a novel domain. Nicolas RH, Goodwin GH; Gene 1996;175:233-240.

[2] Medline: 99198739. The BAH (bromo-adjacent homology) domain: a link between DNA methylation, replication and transcriptional regulation. Callebaut I, Courvalin J-C, Mornon JP; FEBS letts 1999;446:189-193.

30

741. ELM2.

602

ELM2 domain. The ELM2 (Egl-27 and MTA1 homology 2) domain is a small domain of unknown function. Number of members: 10

742. Euk proin. EUKARYOTIC_PORIN The major protein of the outer mitochondrial membrane of eukaryotes is a porin that forms a voltage-dependent anion-selective channel (VDAC) that behaves as a general diffusion pore for small hydrophilic molecules [1 to 4]. The channel adopts an open conformation at low or zero membrane potential and a closed conformation at potentials above 30-40 mV.

This protein contains about 280 amino acids and its sequence is composed of between 12 to 16 beta-strands that span the mitochondrial outer membrane. Yeast contains two members of this family (genes POR1 and POR2); vertebrates have at least three members (genes VDAC1, VDAC2 and VDAC3) [5].

A conserved region located at the C-terminal part of these proteins was selected as a signature pattern.

Consensus pattern[YH]-x(2)-D-[SPCAD]-x-[STA]-x(3)-[TAG]-[KR]-[LIVMF]-[DNSTA]-[DNS]-x(4)-[GSTAN]-[LIVMA]-x-[LIVMY]

[1] Benz R. Biochim. Biophys. Acta 1197:167-196(1994).

[2] Manella C.A. Trends Biochem. Sci. 17:315-320(1992).

[3] Dihanich M. Experientia 46:146-153(1990).

[4] Forte M., Guy H.R., Mannella C.A. J. Bioenerg. Biomembr. 19:341-350(1987).

[5] Sampson M.J., Lovell R.S., Davison D.B., Craigen W.J. Genomics 36:192-196(1996).

743. Glyco hydor 19

Chitinases family 19 signatures

cross-reference(s) CHITINASE_19_1, CHITINASE_19_2

Chitinases (EC 3.2.1.14) [1] are enzymes that catalyze the hydrolysis of the beta-1,4-N-acetyl-D-glucosamine linkages in chitin polymers. From the view point of sequence similarity chitinases belong to either family 18 or 19 in the classification of glycosyl hydrolases [2,E1]. Chitinases of family 19 (also known as classes IA or I and IB or II)

603

are enzymes from plants that function in the defense against fungal and insect pathogens by destroying their chitin-containing cell wall. Class IA/I and IB/II enzymes differ in the presence (IA/I) or absence (IB/II) of a N-terminal chitin-binding domain (see the relevant entry <PDOC00025>). The catalytic domain of these enzymes consist of about 220 to 230 amino acid residues.

Two highly conserved regions were selected as signature patterns, the first one is located in the N-terminal section and contains one of the six cysteines which are conserved in most, if not all, of these chitinases and which is probably involved in a disulfide bond.

Consensus pattern C-x(4,5)-F-Y-[ST]-x(3)-[FY]-[LIVMF]-x-A-x(3)-[YF]-x(2)-F-[GSA]
Consensus pattern [LIVM]-[GSA]-F-x-[STAG](2)-[LIVMFY]-W-[FY]-W-[LIVM]

[1]Flach J., Pilet P.-E., Jolles P. *Experientia* 48:701-716(1992).

[2] Henrissat B. *Biochem. J.* 280:309-316(1991).

744. MBD

Methyl-CpG binding domain

The Methyl-CpG binding domain (MBD) binds to DNA that contains one or more symmetrically methylated CpGs [1]. DNA methylation in animals is associated with alterations in chromatin structure and silencing of gene expression. MBD has negligible non-specific affinity for DNA. In vitro foot-printing with MeCP2 showed the MBD can protect a 12 nucleotide region surrounding a methyl CpG pair [1]. MBDs are found in several Methyl-CpG binding proteins and also DNA demethylase [2]. Number of members: 11

[1]Medline: 94232813. Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. Nan X, Meehan RR, Bird A; *Nucleic Acids Res* 1993;21:4886-4892.

[2]Medline: 99158138. A mammalian protein with specific demethylase activity for mCpG DNA. Bhattacharya SK, Ramchandani S, Cervoni N, Szyf M; *Nature* 1999;397:579-583.

745. Peptidase C1

Eukaryotic thiol (cysteine) proteases active sites

cross-reference(s) THIOLEPROTEASE_CYS; THIOLEPROTEASE_HIS;
THIOLEPROTEASE_ASN

Eukaryotic thiol proteases (EC 3.4.22.-) [1] are a family of proteolytic enzymes which contain an active site cysteine. Catalysis proceeds through a thioester intermediate and is facilitated by a nearby histidine side chain; an asparagine completes the essential catalytic triad. The proteases which are currently known to belong to this family are listed below (references are only provided for recently determined sequences).

- Vertebrate lysosomal cathepsins B (EC 3.4.22.1), H (EC 3.4.22.16), L (EC 3.4.22.15), and S (EC 3.4.22.27) [2].

- Vertebrate lysosomal dipeptidyl peptidase I (EC 3.4.14.1) (also known as cathepsin C) [2].

- Vertebrate calpains (EC 3.4.22.17). Calpains are intracellular calcium-activated thiol protease that contain both a N-terminal catalytic domain and a C-terminal calcium-binding domain.

- Mammalian cathepsin K, which seems involved in osteoclastic bone resorption [3].

- Human cathepsin O [4].

- Bleomycin hydrolase. An enzyme that catalyzes the inactivation of the antitumor drug BLM (a glycopeptide).

- Plant enzymes: barley aleurain (EC 3.4.22.16), EP-B1/B4; kidney bean EP-C1, rice bean SH-EP; kiwi fruit actinidin (EC 3.4.22.14); papaya latex papain (EC 3.4.22.2), chymopapain (EC 3.4.22.6), caricain (EC 3.4.22.30), and proteinase IV (EC 3.4.22.25); pea turgor-responsive protein 15A; pineapple stem bromelain (EC 3.4.22.32); rape COT44; rice oryzain alpha, beta, and gamma; tomato low-temperature induced, Arabidopsis thaliana A494, RD19A and RD21A.

- House-dust mites allergens DerP1 and EurM1.

- Cathepsin B-like proteinases from the worms *Caenorhabditis elegans* (genes gcp-1, cpr-3, cpr-4, cpr-5 and cpr-6), *Schistosoma mansoni* (antigen SM31) and *Japonica* (antigen SJ31), *Haemonchus contortus* (genes AC-1 and AC-2), and *Ostertagia ostertagi* (CP-1 and CP-3).

- Slime mold cysteine proteinases CP1 and CP2.

- Cruzipain from *Trypanosoma cruzi* and *brucei*.

- Throphozoite cysteine proteinase (TCP) from various *Plasmodium* species.

- Proteases from *Leishmania mexicana*, *Theileria annulata* and *Theileria parva*.

605

- Baculoviruses cathepsin-like enzyme (v-cath).
- Drosophila small optic lobes protein (gene sol), a neuronal protein that contains a calpain-like domain.
- Yeast thiol protease BLH1/YCP1/LAP3.
- 5 - Caenorhabditis elegans hypothetical protein C06G4.2, a calpain-like protein.

Two bacterial peptidases are also part of this family:

- Aminopeptidase C from Lactococcus lactis (gene pepC) [5].
- 10 - Thiol protease tpr from Porphyromonas gingivalis.

Three other proteins are structurally related to this family, but may have lost their proteolytic activity.

- 15 - Soybean oil body protein P34. This protein has its active site cysteine replaced by a glycine.

- Rat testin, a sertoli cell secretory protein highly similar to cathepsin L but with the active site cysteine is replaced by a serine. Rat testin should not be confused with mouse testin which is a LIM-domain protein (see <PDOC00382>).

- 20 - Plasmodium falciparum serine-repeat protein (SERA), the major blood stage antigen.

This protein of 111 Kd possesses a C-terminal thiol-protease-like domain [6], but the active site cysteine is replaced by a serine.

The sequences around the three active site residues are well conserved and can be used as signature patterns.

25 Consensus pattern Q-x(3)-[GE]-x-C-[YW]-x(2)-[STAGC]-[STAGCV] [C is the active site residue]

Note the residue in position 4 of the pattern is almost always cysteine; the only exceptions are calpains (Leu), bleomycin hydrolase (Ser) and yeast YCP1 (Ser). Note the residue in position 30 5 of the pattern is always Gly except in papaya protease IV where it is Glu.

Consensus pattern [LIVMGSTAN]-x-H-[GSACE]-[LIVM]-x-[LIVMAT](2)-G-x-[GSADNH]
[H is the active site residue]

Consensus pattern[FYCH]-[WI]-[LIVT]-x-[KRQAG]-N-[ST]-W-x(3)-[FYW]-G-x(2)-G-[LFYW]-[LIVMFYG]-x-[LIVMF] [N is the active site residue]

Note these proteins belong to family C1 (papain-type) and C2 (calpains) in the classification of peptidases [7,E1].

5

[1]Dufour E. Biochimie 70:1335-1342(1988).

[2]Kirschke H., Barrett A.J., Rawlings N.D. Protein Prof. 2:1587-1643(1995).

[3]Shi G.-P., Chapman H.A., Bhairi S.M., Deleeuw C., Reddy V.Y., Weiss S.J. FEBS Lett. 357:129-134(1995).

10 [4]Velasco G., Ferrando A.A., Puente X.S., Sanchez L.M., Lopez-Otin C. J. Biol. Chem. 269:27136-27142(1994).

[5]Chapot-Chartier M.P., Nardi M., Chopin M.C., Chopin A., Gripon J.C. Appl. Environ. Microbiol. 59:330-333(1993).

[6]Higgins D.G., McConnell D.J., Sharp P.M. Nature 340:604-604(1989).

[7]Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:461-486(1994).

746. Peptidase M22

Glycoprotease family signature cross-reference(s) GLYCOPROTEASE

20 Glycoprotease (GCP) (EC 3.4.24.57) [1], or o-sialoglycoprotein endopeptidase, is a metalloprotease secreted by *Pasteurella haemolytica* which specifically cleaves O-sialoglycoproteins such as glycophorin A. The sequence of GCP is highly similar to the following uncharacterized proteins:

- 25 - *Escherichia coli* hypothetical protein ygiD (ORF-X).
- *Bacillus subtilis* hypothetical protein ydiE.
- *Mycobacterium leprae* hypothetical protein U229E.
- *Mycobacterium tuberculosis* hypothetical protein MtCY78.10.
- *Synechocystis* strain PCC 6803 hypothetical protein slr0807.
- 30 - *Methanococcus jannaschii* hypothetical protein MJ1130.
- *Haloarcula marismortui* hypothetical protein in HSH 3'region.
- Yeast hypothetical protein YKR038c.
- Yeast hypothetical protein QRI7.

One of the conserved regions contains two conserved histidines. It is possible that this region is involved in coordinating a metal ion such as zinc.

5 Consensus pattern[**KR**]-[**GSAT**]-x(4)-[**FYWLH**]-[**DQNGK**]-x-P-x-[**LIVMFY**]-x(3)-H-x(2)-[**AG**]-H-[**LIVM**]

Note these proteins belong to family M22 in the classification of peptidases [2,E1].

10 [1]Abdullah K.M., Lo R.Y.C., Mellors A. J. Bacteriol. 173:5597-5603(1991).
[2]Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).

747. SAM. SAM domain (Sterile alpha motif)

15 It has been suggested that SAM is an evolutionarily conserved protein binding domain that is involved in the regulation of numerous developmental processes in diverse eukaryotes. The SAM domain can potentially function as a protein interaction module through its ability to homo- and heterooligomerise with other SAM domains. Number of members: 81

20 [1]Medline: 96100659 SAM: A novel motif in yeast sterile alpha and Drosophila polyhomeotic proteins Ponting CP; Prot Sci 1995;4:1928-1930.

[2]Medline: 97160498 SAM as a protein interaction domain involved in developmental regulation. Shultz J, Ponting CP, Hofmann K, Bork P; Prot Sci 1997;6:249-253.

25 [3]Medline: 99101382 The crystal structure of an Eph receptor SAM domain reveals a mechanism for modular dimerization. Reference Author: Stapleton D, Balan I, Pawson T, Sicheri F; Nat Struct Biol 1999;6:44-49.

748. Tyrosinase signatures cross-reference(s) TYROSINASE_1; TYROSINASE_2

30 Tyrosinase (EC 1.14.18.1) [1] is a copper monooxygenases that catalyzes the hydroxylation of monophenols and the oxidation of o-diphenols to o-quinols. This enzyme, found in prokaryotes as well as in eukaryotes, is involved in the formation of pigments such as melanins and other polyphenolic compounds.

Tyrosinase binds two copper ions (CuA and CuB). Each of the two copper ion has been shown [2] to be bound by three conserved histidines residues. The regions around these copper-binding ligands are well conserved and also shared by some hemocyanins, which are copper-containing oxygen carriers from the hemolymph of many molluscs and arthropods [3,4].

At least two proteins related to tyrosinase are known to exist in mammals:

- TRP-1 (TYRP1) [5], which is responsible for the conversion of 5,6-dihydroxyindole-2-carboxylic acid (DHICA) to indole-5,6-quinone-2-carboxylic acid.
- TRP-2 (TYRP2) [6], which is the melanogenic enzyme DOPAchrome tautomerase (EC 5.3.3.12) that catalyzes the conversion of DOPAchrome to DHICA. TRP-2 differs from tyrosinases and TRP-1 in that it binds two zinc ions instead of copper [7].

Other proteins that belong to this family are:

- Plants polyphenol oxidases (PPO) (EC 1.10.3.1) which catalyze the oxidation of mono- and o-diphenols to o-diquinones [8].
- *Caenorhabditis elegans* hypothetical protein C02C2.1.

Two signature patterns for tyrosinase and related proteins have been derived. The first one contains two of the histidines that bind CuA, and is located in the N-terminal section of tyrosinase. The second pattern contains a histidine that binds CuB, that pattern is located in the central section of the enzyme.

Consensus pattern H-x(4,5)-F-[LIVMFTP]-x-[FW]-H-R-x(2)-[LM]-x(3)-E
[The two H's are copper ligands]

Consensus pattern D-P-x-F-[LIVMFYW]-x(2)-H-x(3)-D [H is a copper ligand]

[1]Lerch K. Prog. Clin. Biol. Res. 256:85-98(1988).

- [2]Jackman M.P., Hajnal A., Lerch K. Biochem. J. 274:707-713(1991).
- [3]Linzen B. Naturwissenschaften 76:206-211(1989).
- [4]Lang W.H., van Holde K.E. Proc. Natl. Acad. Sci. U.S.A. 88:244-248(1991).
- [5]Kobayashi T., Urabe K., Winder A., Jimenez-Cervantes C., Imokawa G., Brewington T., Solano F., Garcia-Borron J.C., Hearing V.J. EMBO J. 13:5818-5825(1994).
- [6]Jackson I.J., Chambers D.M., Tsukamoto K., Copeland N.G., Gilbert D.J., Jenkins N.A., Hearing V. EMBO J. 11:527-535(1992).
- [7]Solano F., Martinez-Liarte J.H., Jimenez-Cervantes C., Garcia-Borron J.C., Lozano J.A. Biochem. Biophys. Res. Commun. 204:1243-1250(1994).
- [8]Cary J.W., Lax A.R., Flurkey W.H. Plant Mol. Biol. 20:245-253(1992).

749. (Mur Ligase) Folylpolyglutamate synthase signatures

Folylpolyglutamate synthase (EC 6.3.2.17) (FPGS) [1] is the enzyme of folate metabolism that catalyzes ATP-dependent addition of glutamate moieties to tetrahydrofolate.

Its sequence is moderately conserved between prokaryotes (gene folC) and eukaryotes. We developed two signature patterns based on the conserved regions which are rich in glycine residues and could play a role in the catalytical activity and/or in substrate binding.

Description of pattern(s) and/or profile(s)

Consensus pattern[LIVMFY]-x-[LIVM]-[STAG]-G-T-[NK]-G-K-x-[ST]-x(7)-[LIVM](2)-x(3)-[GSK]

Consensus pattern[LIVMFY](2)-E-x-G-[LIVM]-[GA]-G-x(2)-D-x-[GST]-x-[LIVM](2)

[1]Shane B., Garrow T., Brenner A., Chen L., Choi Y.J., Hsu J.C., Stover P. Adv. Exp. Med. Biol. 338:629-634(1993).

750. (Peptidase M3) Neutral zinc metallopeptidases, zinc-binding region signature

The majority of zinc-dependent metallopeptidases (with the notable exception of the carboxypeptidases) share a common pattern of primary structure [1,2,3] in the part of their

sequence involved in the binding of zinc, and can be grouped together as a superfamily, known as the metzincins, on the basis of this sequence similarity. They can be classified into a number of distinct families [4,E1] which are listed below along with the proteases which are currently known to belong to these families.

5

Family M1

- Bacterial aminopeptidase N (EC 3.4.11.2) (gene pepN).
- Mammalian aminopeptidase N (EC 3.4.11.2).
- Mammalian glutamyl aminopeptidase (EC 3.4.11.7) (aminopeptidase A). It may play a role in regulating growth and differentiation of early B-lineage cells.
- Yeast aminopeptidase yscII (gene APE2).
- Yeast alanine/arginine aminopeptidase (gene AAP1).
- Yeast hypothetical protein YIL137c.
- Leukotriene A-4 hydrolase (EC 3.3.2.6). This enzyme is responsible for the hydrolysis of an epoxide moiety of LTA-4 to form LTB-4; it has been shown that it binds zinc and is capable of peptidase activity.

10

15

Family M2

- Angiotensin-converting enzyme (EC 3.4.15.1) (dipeptidyl carboxypeptidase I) (ACE) the enzyme responsible for hydrolyzing angiotensin I to angiotensin II. There are two forms of ACE: a testis-specific isozyme and a somatic isozyme which has two active centers.

20

Family M3

- Thimet oligopeptidase (EC 3.4.24.15), a mammalian enzyme involved in the cytoplasmic degradation of small peptides.
- Neurolysin (EC 3.4.24.16) (also known as mitochondrial oligopeptidase M or microsomal endopeptidase).
- Mitochondrial intermediate peptidase precursor (EC 3.4.24.59) (MIP). It is involved the second stage of processing of some proteins imported in the mitochondrion.
- Yeast saccharolysin (EC 3.4.24.37) (proteinase yscD).
- Escherichia coli and related bacteria dipeptidyl carboxypeptidase (EC 3.4.15.5) (gene dcp).
- Escherichia coli and related bacteria oligopeptidase A (EC 3.4.24.70) (gene opdA or prlC).

25

30

- Yeast hypothetical protein YKL134c.

Family M4

- Thermostable thermolysins (EC 3.4.24.27), and related thermolabile neutral proteases (bacillolysins) (EC 3.4.24.28) from various species of *Bacillus*.
- Pseudolysin (EC 3.4.24.26) from *Pseudomonas aeruginosa* (gene *lasB*).
- Extracellular elastase from *Staphylococcus epidermidis*.
- Extracellular protease prt1 from *Erwinia carotovora*.
- Extracellular minor protease smp from *Serratia marcescens*.
- Vibriolysin (EC 3.4.24.25) from various species of *Vibrio*.
- Protease prtA from *Listeria monocytogenes*.
- Extracellular proteinase proA from *Legionella pneumophila*.

Family M5

- Mycolysin (EC 3.4.24.31) from *Streptomyces cacaoi*.

Family M6

- Immune inhibitor A from *Bacillus thuringiensis* (gene *ina*). Ina degrades two classes of insect antibacterial proteins, attacins and cecropins.

Family M7

- *Streptomyces* extracellular small neutral proteases

Family M8

- Leishmanolysin (EC 3.4.24.36) (surface glycoprotein gp63), a cell surface protease from various species of *Leishmania*.

Family M9

- Microbial collagenase (EC 3.4.24.3) from *Clostridium perfringens* and *Vibrio alginolyticus*.

Family M10A

- Serralysin (EC 3.4.24.40), an extracellular metalloprotease from *Serratia*.

- Alkaline metalloproteinase from *Pseudomonas aeruginosa* (gene aprA).
- Secreted proteases A, B, C and G from *Erwinia chrysanthemi*.
- Yeast hypothetical protein YIL108w.

5 Family M10B

- Mammalian extracellular matrix metalloproteinases (known as matrixins) [5]: MMP-1 (EC 3.4.24.7) (interstitial collagenase), MMP-2 (EC 3.4.24.24) (72 Kd gelatinase), MMP-9 (EC 3.4.24.35) (92 Kd gelatinase), MMP-7 (EC 3.4.24.23) (matrylisin), MMP-8 (EC 3.4.24.34) (neutrophil collagenase), MMP-3 (EC 3.4.24.17) (stromelysin-1), MMP-10 (EC 3.4.24.22) (stromelysin-2), and MMP-11 (stromelysin-3), MMP-12 (EC 3.4.24.65) (macrophage metalloelastase).
- Sea urchin hatching enzyme (envelysin) (EC 3.4.24.12). A protease that allows the embryo to digest the protective envelope derived from the egg extracellular matrix.
- Soybean metalloendoproteinase 1.

15 Family M11

- *Chlamydomonas reinhardtii* gamete lytic enzyme (GLE).

20 Family M12A

- Astacin (EC 3.4.24.21), a crayfish endoprotease.
- Meprin A (EC 3.4.24.18), a mammalian kidney and intestinal brush border metalloendopeptidase.
- Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation and which expresses metalloendopeptidase activity. The *Drosophila* homolog of BMP-1 is the dorsal-ventral patterning protein tolloid.
- Blastula protease 10 (BP10) from *Paracentrotus lividus* and the related protein SpAN from *Strongylocentrotus purpuratus*.
- *Caenorhabditis elegans* protein toh-2.
- *Caenorhabditis elegans* hypothetical protein F42A10.8.
- Choriolysins L and H (EC 3.4.24.67) (also known as embryonic hatching proteins LCE and HCE) from the fish *Oryzias latipes*. These proteases participate in the breakdown of the egg envelope, which is derived from the egg extracellular matrix, at the time of hatching.

Family M12B

- Snake venom metalloproteinases [6]. This subfamily mostly groups proteases that act in hemorrhage. Examples are: adamalysin II (EC 3.4.24.46), atrolysin C/D (EC 3.4.24.42), atrolysin E (EC 3.4.24.44), fibrolase (EC 3.4.24.72), trimereylisin I (EC 3.4.25.52) and II (EC 3.4.25.53).
- Mouse cell surface antigen MS2.

Family M13

- Mammalian neprilysin (EC 3.4.24.11) (neutral endopeptidase) (NEP).
- Endothelin-converting enzyme 1 (EC 3.4.24.71) (ECE-1), which process the precursor of endothelin to release the active peptide.
- Kell blood group glycoprotein, a major antigenic protein of erythrocytes. The Kell protein is very probably a zinc endopeptidase.
- Peptidase O from *Lactococcus lactis* (gene pepO).

Family M27

- Clostridial neurotoxins, including tetanus toxin (TeTx) and the various botulinum toxins (BoNT). These toxins are zinc proteases that block neurotransmitter release by proteolytic cleavage of synaptic proteins such as synaptobrevins, syntaxin and SNAP-25 [7,8].

Family M30

- *Staphylococcus hyicus* neutral metalloprotease.

Family M32

- Thermostable carboxypeptidase 1 (EC 3.4.17.19) (carboxypeptidase Taq), an enzyme from *Thermus aquaticus* which is most active at high temperature.

Family M34

- Lethal factor (LF) from *Bacillus anthracis*, one of the three proteins composing the anthrax toxin.

Family M35

- Deuterolysin (EC 3.4.24.39) from *Penicillium citrinum* and related proteases from various species of *Aspergillus*.

5 Family M36

- Extracellular elastinolytic metalloproteinases from *Aspergillus*.

10 From the tertiary structure of thermolysin, the position of the residues acting as zinc ligands and those involved in the catalytic activity are known. Two of the zinc ligands are histidines which are very close together in the sequence; C-terminal to the first histidine is a glutamic acid residue which acts as a nucleophile and promotes the attack of a water molecule on the carbonyl carbon of the substrate. A signature pattern which includes the two histidine and the glutamic acid residues is sufficient to detect this superfamily of proteins.

Description of pattern(s) and/or profile(s)

Consensus pattern[GSTALIVN]-x(2)-H-E-[LIVMFYW]-{DEHRKP}-H-x-[LIVMFYWGSPQ] [The two H's are zinc ligands] [E is the active site residue]

20 Sequences known to belong to this class detected by the patternALL, except for members of families M5, M7 and M11.

Other sequence(s) detected in SWISS-PROT55; including *Neurospora crassa* conidiation-specific protein 13 which could be a zinc-protease.

25 [1]Jongeneel C.V., Bouvier J., Bairoch A.
FEBS Lett. 242:211-214(1989).

[2]Murphy G.J.P., Murphy G., Reynolds J.J.
FEBS Lett. 289:4-7(1991).

[3]Bode W., Grams F., Reinemer P., Gomis-Rueth F.-X., Baumann U., McKay
30 D.B., Stoecker W.
Zoology 99:237-246(1996).

[4]Rawlings N.D., Barrett A.J.
Meth. Enzymol. 248:183-228(1995).

[5]Woessner J. Jr.

FASEB J. 5:2145-2154(1991).

[6]Hite L.A., Fox J.W., Bjarnason J.B.

[7]Montecucco C., Schiavo G.

5 Trends Biochem. Sci. 18:324-327(1993).

[8]Niemann H., Blasi J., Jahn R.

Trends Cell Biol. 4:179-185(1994).

10 751. PseudoU_synt_1

tRNA pseudouridine synthase is involved in the formation of pseudouridine at the anticodon stem and loop of transfer-RNAs Pseudouridine is an isomer of uridine (5-(beta-D-ribofuranosyl) uracil, and is the most abundant modified nucleoside found in all cellular RNAs. The TruA-like proteins also exhibit a conserved sequence with a strictly conserved aspartic acid, likely involved in catalysis. Number of members: 25

[1]Medline: 98254513. Transfer RNA-pseudouridine synthetase Pus1 of *Saccharomyces cerevisiae* contains one atom of zinc essential for its native conformation and tRNA recognition. Arluison V, Hountondji C, Robert B, Grosjean H; *Biochemistry* 1998;37:7268-7276.

752. EPSP synthase signatures

EPSP synthase (3-phosphoshikimate 1-carboxyvinyltransferase) (EC 2.5.1.19) catalyzes the sixth step in the biosynthesis from chorismate of the aromatic amino acids (the shikimate pathway) in bacteria (gene *aroA*), plants and fungi (where it is part of a multifunctional enzyme which catalyzes five consecutive steps in this pathway) [1]. EPSP synthase has been extensively studied as it is the target of the potent herbicide glyphosate which inhibits the enzyme.

The sequence of EPSP from various biological sources shows that the structure of the enzyme has been well conserved throughout evolution. Two conserved regions were selected as signature patterns. The first pattern corresponds to a region that is part of the active site and

616

which is also important for the resistance to glyphosate [2]. The second pattern is located in the C-terminal part of the protein and contains a conserved lysine which seems to be important for the activity of the enzyme.

5 Description of pattern(s) and/or profile(s)

Consensus pattern[LIVM]-x(2)-[GN]-N-[SA]-G-T-[STA]-x-R-x-[LIVMY]-x-[GSTA]
Consensus pattern[KR]-x-[KH]-E-[CST]-[DNE]-R-[LIVM]-x-[STA]-[LIVMC]-x(2)-[EN]-
[LIVMF]-x-[KRA]-[LIVMF]-G

10

[1]Stallings W.C., Abdel-Megid S.S., Lim L.W., Shieh H.-S., Dayringer H.E., Leimgruber N.K., Stegeman R.A., Anderson K.S., Sikorski J.A., Padgett S.R., Kishore G.M. Proc. Natl. Acad. Sci. U.S.A. 88:5046-5050(1991).
[2]Padgett S.R., Re D.B., Gaser C.S., Eicholtz D.A., Frazier R.B., Hironaka C.M., Levine E.B., Shah D.M., Fraley R.T., Kishore G.M. J. Biol. Chem. 266:22364-22369(1991).

15

753. Glyco_hydro_18

Glycosyl hydrolases family 18. Number of members: 173

[1]Medline: 95219379. Crystal structure of a bacterial chitinase at 2.3 Å resolution. Perrakis A, Tews I, Dauter Z, Oppenheim AB, Chet I, Wilson KS, Vorgias CE; Structure 1994;2:1169-1180.

20

25 754. Esterase

Putative esterase

This family contains Esterase D Swiss:P10768. However it is not clear if all members of the family have the same function. This family is possibly related to the COesterase family.

Number of members: 36

30

755. (HMA) Heavy-metal-associated domain

A conserved domain of about 30 amino acid residues has been found [1] in a number of proteins that transport or detoxify heavy metals. This domain contains two conserved cysteines that could be involved in the binding of these metals. The domain has been termed Heavy-Metal-Associated (HMA). It has been found in:

- A variety of cation transport ATPases (E1-E2 ATPases) (see <PDOC00139>). The human copper ATPases ATP7A and ATP7B which are respectively involved in Menke's and Wilson's diseases. ATP7A and ATP7B both contain 6 tandem copies of the HMA domain. The copper ATPases CCC2 from budding yeast, copA from *Enterococcus faecalis* and synA from *Synechococcus* contain one copy of the HMA domain. The cadmium ATPases cadA from *Bacillus firmus* and from plasmid pI258 from *Staphylococcus aureus* also contain a single HMA domain, while a chromosomal *Staphylococcus aureus* cadA contains two copies. Other, less characterized ATPases that contain the HMA domain are: fixI from *Rhizobium meliloti*, pacS from *Synechococcus* strain PCC 7942), *Mycobacterium leprae* ctpA and ctpB and *Escherichia coli* hypothetical protein yhhO. In all these ATPases the HMA domain(s) are located in the N-terminal section.
- Mercuric reductase (EC 1.16.1.1) (gene merA) which is generally encoded by plasmids carried by mercury-resistant Gram-negative bacteria. Mercuric reductase is a class-1 pyridine nucleotide-disulphide oxidoreductase (see <PDOC00073>). There is generally one HMA domain (with the exception of a chromosomal merA from *Bacillus* strain RC607 which has two) in the N-terminal part of merA.
- Mercuric transport protein periplasmic component (gene merP), also encoded by plasmids carried by mercury-resistant Gram-negative bacteria. It seems to be a mercury scavenger that specifically binds to one Hg(2+) ion and which passes it to the mercuric reductase via the merT protein. The N-terminal half of merP is a HMA domain.
- *Helicobacter pylori* copper-binding protein copP.
- Yeast protein ATX1 [2], which could act in the transport and/or partitioning of copper.

The consensus pattern for HMA spans the complete domain.

Description of pattern(s) and/or profile(s)

Consensus pattern[LIVN]-x(2)-[LIVMFA]-x-C-x-[STAGCDNH]-C-x(3)-[LIVFG]-x(3)-
[LIV]-x(9,11)-[IVA]-x-[LVFYS] [The two C's probably bind metals]

[1]Bull P.C., Cox D.W. Trends Genet. 10:246-252(1994).

5 [2]Lin S.-J., Culotta V.L. Proc. Natl. Acad. Sci. U.S.A. 92:3784-3788(1995).

756. (Peptidase M10) Matrixins cysteine switch

PROSITE cross-reference(s): CYSTEINE_SWITCH

Mammalian extracellular matrix metalloproteinases (EC 3.4.24.-), also known as matrixins

10 [1] (see <PDOC00129>), are zinc-dependent enzymes. They are secreted by cells in an
inactive form (zymogen) that differs from the mature enzyme by the presence of an N-
terminal propeptide. A highly conserved octapeptide is found two residues downstream of
the C-terminal end of the propeptide. This region has been shown to be involved in
autoinhibition of matrixins [2,3]; a cysteine within the octapeptide chelates the active site
15 zinc ion, thus inhibiting the enzyme. This region has been called the 'cysteine switch' or
'autoinhibitor region'.

A cysteine switch has been found in the following zinc proteases:

- MMP-1 (EC 3.4.24.7) (interstitial collagenase).
- 20 - MMP-2 (EC 3.4.24.24) (72 Kd gelatinase).
- MMP-3 (EC 3.4.24.17) (stromelysin-1).
- MMP-7 (EC 3.4.24.23) (matrilysin).
- MMP-8 (EC 3.4.24.34) (neutrophil collagenase).
- MMP-9 (EC 3.4.24.35) (92 Kd gelatinase).
- 25 - MMP-10 (EC 3.4.24.22) (stromelysin-2).
- MMP-11 (EC 3.4.24.-) (stromelysin-3).
- MMP-12 (EC 3.4.24.65) (macrophage metalloelastase).
- MMP-13 (EC 3.4.24.-) (collagenase 3).
- MMP-14 (EC 3.4.24.-) (membrane-type matrix metalloproteinase 1).
- 30 - MMP-15 (EC 3.4.24.-) (membrane-type matrix metalloproteinase 2).
- MMP-16 (EC 3.4.24.-) (membrane-type matrix metalloproteinase 3).
- Sea urchin hatching enzyme (EC 3.4.24.12) (envelysin) [4].
- Chlamydomonas reinhardtii gamete lytic enzyme (GLE) [5].

Description of pattern(s) and/or profile(s)

Consensus pattern P-R-C-[GN]-x-P-[DR]-[LIVSAPKQ] [C chelates the zinc ion]

- 5 [1]Woessner J. Jr. FASEB J. 5:2145-2154(1991).
 [2]Sanchez-Lopez R., Nicholson R., Gesnel M.C., Matrisian L.M., Breathnach R. J. Biol. Chem. 263:11892-11899(1988).
 [3]Park A.J., Matrisian L.M., Kells A.F., Pearson R., Yuan Z., Navre M. J. Biol. Chem. 266:1584-1590(1991).
 10 [4]Lepage T., Gache C. EMBO J. 9:3003-3012(1990).
 [5]Kinoshita T., Fukuzawa H., Shimada T., Saito T., Matsuda Y. Proc. Natl. Acad. Sci. U.S.A. 89:4693-4697(1992).

757. (Peptidase S8) Serine proteases, subtilase family, active sites

PROSITE cross-reference(s): PS00136; SUBTILASE_ASP, PS00137; SUBTILASE_HIS, PS00138; SUBTILASE_SER

Subtilases [1,2] are an extensive family of serine proteases whose catalytic activity is provided by a charge relay system similar to that of the trypsin family of serine proteases but which evolved by independent convergent evolution. The sequence around the residues involved in the catalytic triad (aspartic acid, serine and histidine) are completely different from that of the analogous residues in the trypsin serine proteases and can be used as signatures specific to that category of proteases.

The subtilase family currently includes the following proteases:

- 25 - Subtilisins (EC 3.4.21.62), these alkaline proteases from various *Bacillus* species have been the target of numerous studies in the past thirty years.
- Alkaline elastase YaB from *Bacillus* sp. (gene ale).
 - Alkaline serine exoprotease A from *Vibrio alginolyticus* (gene proA).
 - Aqualysin I from *Thermus aquaticus* (gene pstI).
- 30 - AspA from *Aeromonas salmonicida*.
- Bacillopeptidase F (esterase) from *Bacillus subtilis* (gene bpf).
 - C5A peptidase from *Streptococcus pyogenes* (gene scpA).
 - Cell envelope-located proteases PI, PII, and PIII from *Lactococcus lactis*.

620

- Extracellular serine protease from *Serratia marcescens*.
 - Extracellular protease from *Xanthomonas campestris*.
 - Intracellular serine protease (ISP) from various *Bacillus*.
 - Minor extracellular serine protease epr from *Bacillus subtilis* (gene epr).
 - 5 - Minor extracellular serine protease vpr from *Bacillus subtilis* (gene vpr).
 - Nisin leader peptide processing protease nisP from *Lactococcus lactis*.
 - Serotype-specific antigene 1 from *Pasteurella haemolytica* (gene ssa1).
 - Thermitase (EC 3.4.21.66) from *Thermoactinomyces vulgaris*.
 - Calcium-dependent protease from *Anabaena variabilis* (gene prcA).
 - 10 - Halolysin from halophilic bacteria sp. 172p1 (gene hly).
 - Alkaline extracellular protease (AEP) from *Yarrowia lipolytica* (gene xpr2).
 - Alkaline proteinase from *Cephalosporium acremonium* (gene alp).
 - Cerevisin (EC 3.4.21.48) (vacuolar protease B) from yeast (gene PRB1).
 - Cuticle-degrading protease (pr1) from *Metarhizium anisopliae*.
 - 15 - KEX-1 protease from *Kluyveromyces lactis*.
 - Kexin (EC 3.4.21.61) from yeast (gene KEX-2).
 - Oryzin (EC 3.4.21.63) (alkaline proteinase) from *Aspergillus* (gene alp).
 - Proteinase K (EC 3.4.21.64) from *Tritirachium album* (gene proK).
 - Proteinase R from *Tritirachium album* (gene proR).
 - 20 - Proteinase T from *Tritirachium album* (gene proT).
 - Subtilisin-like protease III from yeast (gene YSP3).
 - Thermomycin (EC 3.4.21.65) from *Malbranchea sulfurea*.
 - Furin (EC 3.4.21.85), neuroendocrine convertases 1 to 3 (NEC-1 to -3) and PACE4
- protease from mammals, other vertebrates, and invertebrates. These proteases are involved
- 25 in the processing of hormone precursors at sites comprised of pairs of basic amino acid residues [3].
- Tripeptidyl-peptidase II (EC 3.4.14.10) (tripeptidyl aminopeptidase) from Human.
 - Prestalk-specific proteins tagB and tagC from slime mold [4]. Both proteins consist of two domains: a N-terminal subtilase catalytic domain and a C-terminal ABC transporter domain
 - 30 (see <PDOC00185>).

Description of pattern(s) and/or profile(s)

621

Consensus pattern[STAIV]-x-[LIVMF]-[LIVM]-D-[DSTA]-G-[LIVMFC]-x(2,3)-[DNH] [D is the active site residue]

Consensus patternH-G-[STM]-x-[VIC]-[STAGC]-[GS]-x-[LIVMA]-[STAGCLV]-[SAGM] [H is the active site residue]

5 Consensus patternG-T-S-x-[SA]-x-P-x(2)-[STAVC]-[AG] [S is the active site residue]

Note if a protein includes at least two of the three active site signatures, the probability of it being a serine protease from the subtilase family is 100%

Note these proteins belong to family S8 in the classification of peptidases [5,E1].

10

[1]Siezen R.J., de Vos W.M., Leunissen J.A.M., Dijkstra B.W. Protein Eng. 4:719-737(1991).

[2]Siezen R.J. (In) Proceeding subtilisin symposium, Hamburg, (1992).

[3]Barr P.J. Cell 66:1-3(1991).

15 [4]Shaulsky G., Kuspa A., Loomis W.F.; Genes Dev. 9:1111-1122(1995).

[5]Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

758. (SSB) Single-strand binding protein family signatures

PROSITE cross-reference(s): PS00735; SSB_1,PS00736; SSB_2

The Escherichia coli single-strand binding protein [1] (gene ssb), also known as the helix-destabilizing protein, is a protein of 177 amino acids. It binds tightly, as a homotetramer, to single-stranded DNA (ss-DNA) and plays an important role in DNA replication, recombination and repair.

25

Closely related variants of SSB are encoded in the genome of a variety of large self-transmissible plasmids. SSB has also been characterized in bacteria such as Proteus mirabilis or Serratia marcescens.

30 Eukaryotic mitochondrial proteins that bind ss-DNA and are probably involved in mitochondrial DNA replication are structurally and evolutionary related to prokaryotic SSB. Proteins currently known to belong to this subfamily are listed below [2].

- Mammalian protein Mt-SSB (P16).

- *Xenopus* Mt-SSBs and Mt-SSBr.
- *Drosophila* MtSSB.
- Yeast protein RIM1.

5 Two signature patterns have been developed for these proteins. The first is a conserved region in the N-terminal section of the SSB's. The second is a centrally located region which, in *Escherichia coli* SSB, is known to be involved in the binding of DNA.

Description of pattern(s) and/or profile(s)

10 Consensus pattern[LIVMF]-[NST]-[KRT]-[LIVM]-x-[LIVMF](2)-G-[NHRK]-[LIVM]-[GST]-x-[DET]

Consensus patternT-x-W-[HY]-[RNS]-[LIVM]-x-[LIVMF]-[FY]-[NGKR]

[1]Meyer R.R., Laine P.S. Microbiol. Rev. 54:342-380(1990).

15 [2]Stroumbakis N.D., Li Z., Tolias P.P. Gene 143:171-177(1994).

759. KDPG and KHG aldolases active site signatures

PROSITE cross-reference(s): PS00159; ALDOLASE_KDPG_KHG_1, PS00160;

ALDOLASE_KDPG_KHG_2

20 4-hydroxy-2-oxoglutarate aldolase (EC 4.1.3.16) (KHG-aldolase) catalyzes the interconversion of 4-hydroxy-2-oxoglutarate into pyruvate and glyoxylate. Phospho-2-dehydro-3-deoxygluconate aldolase (EC 4.1.2.14) (KDPG-aldolase) catalyzes the interconversion of 6-phospho-2-dehydro-3-deoxy-D-gluconate into pyruvate and
25 glyceraldehyde 3-phosphate.

These two enzymes are structurally and functionally related [1]. They are both homotrimeric proteins of approximately 220 amino-acid residues. They are class I aldolases whose catalytic mechanism involves the formation of a Schiff-base intermediate between the substrate and
30 the epsilon-amino group of a lysine residue. In both enzymes, an arginine is required for catalytic activity.

623

Two signature patterns were developed for these enzymes. The first one contains the active site arginine and the second, the lysine involved in the Schiff-base formation.

Description of pattern(s) and/or profile(s)

- 5 Consensus pattern G-[LIVM]-x(3)-E-[LIV]-T-[LF]-R [R is the active site residue]
Consensus pattern G-x(3)-[LIVMF]-K-[LF]-F-P-[SA]-x(3)-G [K is involved in Schiff-base formation]

[1] Vlahos C J., Dekker E.E. J. Biol. Chem. 263:11683-11691(1988).

10

760. AP endonucleases family 1 signatures. PROSITE cross-reference(s): PS00726;
AP_NUCLEASE_F1_1, PS00727; AP_NUCLEASE_F1_2, PS00728;
AP_NUCLEASE_F1_3

Sequence

- 5 DNA damaging agents such as the antitumor drugs bleomycin and neocarzinostatin or those that generate oxygen radicals produce a variety of lesions in DNA. Amongst these is base-loss which forms apurinic/aprimidinic (AP) sites or strand breaks with atypical 3'termini. DNA repair at the AP sites is initiated by specific endonuclease cleavage of the phosphodiester backbone. Such endonucleases are also generally capable of removing blocking groups from the 3'terminus of DNA strand breaks.

20

AP endonucleases can be classified into two families on the basis of sequence similarity. Family 1 groups the enzymes listed below [1].

- 25 - Escherichia coli exonuclease III (EC 3.1.11.2) (gene xthA).
- Streptococcus pneumoniae and Bacillus subtilis exonuclease A (gene exoA).
- Mammalian AP endonuclease 1 (AP1) (EC 4.2.99.18).
- Drosophila recombination repair protein 1 (gene Rrp1).
- Arabidopsis thaliana apurinic endonuclease-redox protein (gene arp).

30

Except for Rrp1 and arp, these enzymes are proteins of about 300 amino-acid residues. Rrp1 and arp both contain additional and unrelated sequences in their N-terminal section (about 400 residues for Rrp1 and 270 for arp).

Three signature patterns were developed for this family of enzymes. The patterns are based on the most conserved regions. The first pattern contains a glutamate which has been shown [2], in the Escherichia coli enzyme to bind a divalent metal ion such as magnesium or manganese

Consensus pattern[APF]-D-[LIVMF](2)-x-[LIVM]-Q-E-x-K [E binds a divalent metal ion]
 Consensus patternD-[ST]-[FY]-R-[KH]-x(7,8)-[FYW]-[ST]-[FYW](2)
 Consensus patternN-x-G-x-R-[LIVM]-D-[LIVMFYH]-x-[LV]-x-S

[1] Barzilay G., Hickson I.S. BioEssays 17:713-719(1995).

[2] Mol C.D., Kuo C.-F., Thayer M.M., Cunningham R.P., Tainer J.A. Nature 374:381-386(1995).

761. (ER)Enhancer of rudimentary signature, PROSITE cross-reference(s): PS01290; ER

The Drosophila protein 'enhancer of rudimentary' (gene (e(r)) is a small protein of 104 residues whose function is not yet clear. From an evolutionary point of view, it is highly conserved [1] and has been found to exist in probably all multicellular eukaryotic organisms. It has been proposed that this protein plays a role in the cell cycle.

A conserved region in the central part of the protein was selected as as signaure pattern.

Consensus patternY-D-I-[SA]-x-L-[FY]-x-F-[IV]-D-x(3)-D-[LIV]-S

[1] Gelsthorpe M., Pulumati M., McCallum C., Dang-Vu K., Tsubota S.I. Gene 186:189-195(1997).

762. (ETF alpha) Electron transfer flavoprotein alpha-subunit signature, PROSITE cross-reference(s): PS00696; ETF_ALPHA

The electron transfer flavoprotein (ETF) [1,2] serves as a specific electron acceptor for various mitochondrial dehydrogenases. ETF transfers electrons to the main respiratory

chain via ETF-ubiquinone oxidoreductase. ETF is an heterodimer that consist of an alpha and a beta subunit and which bind one molecule of FAD per dimer. A similar system also exists in some bacteria.

- 5 The alpha subunit of ETF is a protein of about 32 Kd which is structurally related to the bacterial nitrogen fixation protein fixB which could play a role in a redox process and feed electrons to ferredoxin.

Other related proteins are:

10

- Escherichia coli hypothetical protein ydiR.
- Escherichia coli hypothetical protein ygcQ.

A highly conserved region which is located in the C-terminal section was selected as a signature pattern for these proteins.

15

Consensus pattern [LI]-Y-[LIVM]-[AT]-x-G-[IV]-[SD]-G-x-[IV]-Q-H-x(2)-G-x(6)-[IV]-x-A-[IV]-N

20

- [1] Finocchiaro G., Ikeda Y., Ito M., Tanaka K. Prog. Clin. Biol. Res. 321:637-652(1990).
- [2] Tsai M.H., Saier M.H. Jr. Res. Microbiol. 146:397-404(1995).

763. (lectin c) C-type lectin domain signature and profile

PROSITE cross-reference(s): PS00615; C_TYPE_LLECTIN_1, PS50041;

25 C_TYPE_LLECTIN_2

30

A number of different families of proteins share a conserved domain which was first characterized in some animal lectins and which seem to function as a calcium-dependent carbohydrate-recognition domain [1,2,3]. This domain, which is known as the C-type lectin domain (CTL) or as the carbohydrate-recognition domain (CRD), consists of about 110 to 130 residues. There are four cysteines which are perfectly conserved and involved in two disulfide bonds. A schematic representation of the CTL domain is shown below.

```

      +-----+
      |   |
xcxxxxcxxxxxxxxCxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxCxxxxWxCxxxxCx
|   |   | *****|*
5  +----+   +-----+

```

'C': conserved cysteine involved in a disulfide bond.

'c': optional cysteine involved in a disulfide bond.

'*': position of the pattern.

The categories of proteins, in which the CTL domain has been found, are listed below.

Type-II membrane proteins where the CTL domain is located at the C-terminal extremity of the proteins:

- Asialoglycoprotein receptors (ASGPR) (also known as hepatic lectins) [4]. The ASGPR's mediate the endocytosis of plasma glycoproteins to which the terminal sialic acid residue in their carbohydrate moieties has been removed.
- Low affinity immunoglobulin epsilon Fc receptor (lymphocyte IgE receptor), which plays an essential role in the regulation of IgE production and in the differentiation of B cells.
- Kupffer cell receptor. A receptor with an affinity for galactose and fucose, that could be involved in endocytosis.
- A number of proteins expressed on the surface of natural killer T-cells: NKG2, NKR-P1, YE1/88 (Ly-49), CD69 and on B-cells: CD72, LyB-2. The CTL- domain in these proteins is distantly related to other CTL-domains; it is unclear whether they are likely to bind carbohydrates.

Proteins that consist of an N-terminal collagenous domain followed by a CTL- domain [5], these proteins are sometimes called 'collectins':

- Pulmonary surfactant-associated protein A (SP-A). SP-A is a calcium-dependent protein that binds to surfactant phospholipids and contributes to lower the surface tension at the air-liquid interface in the alveoli of the

mammalian lung.

- Pulmonary surfactant-associated protein D (SP-D).
- Conglutinin, a calcium-dependent lectin-like protein which binds to a yeast cell wall extract and to immune complexes through the complement component (iC3b).
- Mannan-binding proteins (MBP) (also known as mannose-binding proteins). MBP's bind mannose and N-acetyl-D-glucosamine in a calcium-dependent manner.
- Bovine collectin-43 (CL-43).

Selectins (or LEC-CAM) [6,7]. Selectins are cell adhesion molecules implicated in the interaction of leukocytes with platelets or vascular endothelium. Structurally, selectins consist of a long extracellular domain, followed by a transmembrane region and a short cytoplasmic domain. The extracellular domain is itself composed of a CTL-domain, followed by an EGF-like domain and a variable number of SCR/Sushi repeats. Known selectins are:

- Lymph node homing receptor (also known as L-selectin, leukocyte adhesion molecule-1, (LAM-1), leu-8, gp90-mel, or LECAM-1)
- Endothelial leukocyte adhesion molecule 1 (ELAM-1, E-selectin or LECAM-2). The ligand recognized by ELAM-1 is sialyl-Lewis x.
- Granule membrane protein 140 (GMP-140, P-selectin, PADGEM, CD62, or LECAM-3). The ligand recognized by GMP-140 is Lewis x.

Large proteoglycans that contain a CTL-domain followed by one copy of a SCR/ Sushi repeat, in their C-terminal section:

- Aggrecan (cartilage-specific proteoglycan core protein). This proteoglycan is a major component of the extracellular matrix of cartilaginous tissues where it has a role in the resistance to compression.
- Brevican.
- Neurocan.
- Versican (large fibroblast proteoglycan), a large chondroitin sulfate

proteoglycan that may play a role in intercellular signalling.

In addition to the CTL and Sushi domains, these proteins also contain, in their N-terminal domain, an Ig-like V-type region, two or four link domains (see <PDOC00955>) and up to two EGF-like repeats.

Two type-I membrane proteins:

- Mannose receptor from macrophages. This protein mediates the endocytosis of glycoproteins by macrophages in several recognition and uptake processes. Its extracellular section consists of a fibronectin type II domain followed by eight tandem repeats of the CTL domain.
- 180 Kd secretory phospholipase A2 receptor (PLA2-R). A protein whose structure is highly similar to that of the mannose receptor.
- DEC-205 receptor. This protein is used by dendritic cells and thymic epithelial cells to capture and endocytose diverse carbohydrate-binding antigens and direct them to antigen-processing cellular compartments. DEC-205 extracellular section consists of a fibronectin type II domain followed by ten tandem repeats of the CTL domain.
- Silk moth hemocytin, an humoral lectin which is involved in a self-defence mechanism. It is composed of 2 FA58C domains (see <PDOC00988>), a CTL domain, 2 VWFC domains (see <PDOC00928>), and a CTCK (see <PDOC00912>).

Various other proteins that uniquely consist of a CTL domain:

- Invertebrate soluble galactose-binding lectins. A category to which belong a humoral lectin from a flesh fly; echinoidin, a lectin from the coelomic fluid of a sea urchin; BRA-2 and BRA-3, two lectins from the coelomic fluid of a barnacle, a lectin from the tunicate *Polyandrocarpa misakiensis* and a newt oviduct lectin. The physiological importance of these lectins is not yet known but they may play an important role in defense mechanisms.
- Pancreatic stone protein (PSP) (also known as pancreatic thread protein (PTP), or reg), a protein that might act as an inhibitor of spontaneous

calcium carbonate precipitation.

- Pancreatitis associated protein (PAP), a protein that might be involved in the control of bacterial proliferation.
- Tetranectin, a plasma protein that binds to plasminogen and to isolated
5 kringle 4.
- Eosinophil granule major basic protein (MBP), a cytotoxic protein.
- A galactose specific lectin from a rattlesnake.
- Two subunits of a coagulation factor IX/factor X-binding protein (IX/X-bp),
10 a snake venom anticoagulant protein which binds with factors IX and X in the presence of calcium.
- Two subunits of a phospholipase A2 inhibitor from the plasma of a snake (PLI-A and PLI-B).
- A lipopolysaccharide-binding protein (LPS-BP) from the hemolymph of a cockroach [8].
- Sea raven antifreeze protein (AFP) [9].

As a signature pattern for this domain, the C-terminal region with its three conserved cysteines was selected.

Consensus pattern C-[LIVMFYATG]-x(5,12)-[WL]-x-[DNSR]-x(2)-C-x(5,6)-
20 [FYWLIVSTA]-[LIVMSTA]-C [The three C's are involved in disulfide bonds]

Note all CTL domains have five Trp residues before the second Cys,
25 with the exception of tunicate lectin and cockroach LPS-BP which have Leu.

Note this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the
30 pattern, you should use it if you have access to the necessary software tools to do so.

[1] Drickamer K. J. Biol. Chem. 263:9557-9560(1988).

[2] Drickamer K. Prog. Nucleic Acid Res. Mol. Biol. 45:207-232(1993).

- [3] Drickamer K. Curr. Opin. Struct. Biol. 3:393-400(1993).
- [4] Spiess M. Biochemistry 29:10009-10018(1990).
- [5] Weis W.I., Kahn R., Fourme R., Drickamer K., Hendrickson W.A. Science 254:1608-1615(1991).
- 5 [6] Siegelman M. Curr. Biol. 1:125-128(1991).
- [7] Lasky L.A. Science 238:964-969(1992).
- [8] Jomori T., Natori S. J. Biol. Chem. 266:13318-13323(1991).
- [9] Ng N.F.L., Hew C.-L. J. Biol. Chem. 267:16069-16075(1992).

10 764. (SRCR) Speract receptor repeated domain signature
 PROSITE cross-reference(s): PS00420; SPERACT_RECEPTOR,

The receptor for the sea urchin egg peptide speract is a transmembrane glycoprotein of 500 amino acid residues [1]. Structurally it consists of a large extracellular domain of 450 residues, followed by a transmembrane region and a small cytoplasmic domain of 12 amino acids. The extracellular domain contains four repeats of a 115 amino acids domain. There are 17 positions that are perfectly conserved in the four repeats, among them are six cysteines, six glycines, and three glutamates.

20 Such a domain is also found, once, in the C-terminal section of mammalian macrophage scavenger receptor type I [2], a membrane glycoproteins implicated in the pathologic deposition of cholesterol in arterial walls during atherogenesis.

25 The signature pattern that was derived spans part of the N-terminal section of the domain and contains 8 of the 17 conserved residues.

Consensus pattern G-x(5)-G-x(2)-E-x(6)-W-G-x(2)-C-x(3)-[FYW]-x(8)-C-x(3)-G

- [1] Dangott J.J., Jordan J.E., Bellet R.A., Garbers D.L. Proc. Natl. Acad. Sci. U.S.A. 86:2128-2132(1989).
- 30 [2] Freeman M., Ashkenas J., Rees D.J., Kingsley D.M., Copeland N.G., Jenkins N.A., Krieger M. Proc. Natl. Acad. Sci. U.S.A. 87:8810-8814(1990).

765. Bac_surface_Ag

Bacterial surface antigen

This entry includes the following surface antigens; D15 antigen from H.influenzae, OMA87 from P.multocida, OMP85 from N.meningitidis and N.gonorrhoeae. Number of members:

5 14

[1]Medline: 95255676. The sequencing of the 80-kDa D15 protective surface antigen of Haemophilus influenzae. Flack FS, Loosmore S, Chong P, Thomas WR; Gene 1995;156:97-99.

10 [2] Medline: 96333354. Cloning, sequencing, expression, and protective capacity of the oma87 gene encoding the Pasteurella multocida 87-kilodalton outer membrane antigen. Ruffolo CG, Adler B; Infect Immun 1996;64:3161-3167.

766. BRCA1 C Terminus (BRCT) domain

5 The BRCT domain is found predominantly in proteins involved in cell cycle checkpoint functions responsive to DNA damage. It has been suggested that the Retinoblastoma protein contains a divergent BRCT domain, this has not been included in this family. The BRCT domain of XRCC1 forms a homodimer in the crystal structure Medline:99016060. This suggests that pairs of BRCT domains

20 associate as homo- or heterodimers. Number of members: 131

[1] Medline: 96259550. BRCA1 protein products ...Functional motifs... Koonin EV, Altschul SF, Bork P; Nature Genet 1996;13:266-268.

25 [2] Medline: 97153217. From BRCA1 to RAP1: A widespread BRCT module closely associated with DNA repair Callebaut I, Mornon JP; Febs lett 1997;400:25-30.

[3] Medline: 97186552. A superfamily of conserved domains in DNA damage responsive cell cycle checkpoint proteins Bork P, Hofmann K, Bucher P, Neuwald AF, Altschul SF, Koonin EV; Faseb J 1997;11:68-76.

30 [4] Medline: 97402527. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ; Nucleic Acids Res 1997;25:3389-3402.

[5] Medline: 99016060. Structure of an XRCC1 BRCT domain: a new protein-protein interaction module. Zhang X, Morera S, Bates PA, Whitehead PC, Coffey AI, Hainbucher K, Nash RA, Sternberg MJ, Lindahl T, Freemont PS;

5 767. Kappa casein

Kappa-casein is a mammalian milk protein involved in a number of important physiological processes. In the gut, the ingested protein is split into an insoluble peptide (para kappa-casein) and a soluble hydrophilic glycopeptide (caseinomacropeptide). Caseinomacropeptide is responsible for increased efficiency of digestion, prevention of neonate hypersensitivity to ingested proteins, and inhibition of gastric pathogens. Number of members: 56

[1] Medline: 98072500. Nucleotide sequence evolution at the kappa-casein locus: evidence for positive selection within the family Bovidae. Ward TJ, Honeycutt RL, Derr JN; Genetics 1997;147:1863-1872.

15 768. Chitinases family 18 active site

PROSITE cross-reference(s) CHITINASE_18

Chitinases (EC 3.2.1.14) [1] are enzymes that catalyze the hydrolysis of the beta-1,4-N-acetyl-D-glucosamine linkages in chitin polymers. From the view point of sequence similarity chitinases belong to either family 18 or 19 in the classification of glycosyl hydrolases [2,E1]. Chitinases of family 18 (also known as classes III or V) groups a variety of proteins:

a) Chitinases from:

- 25 - Prokaryotes such as Alteromonas, Bacillus, Serratia, Streptomyces, etc.
- Plants such as Arabidopsis, cucumber, bean, tobacco, etc.
- Fungi such as Aphanocladium, Rhizopus, Saccharomyces, etc.
- Nematode (Brugia malayi).
- Insects (Manduca sexta).
- 30 - Baculoviruses (Autographa Californica Nuclear Polyhedrosis virus).

b) Other proteins:

633

- Hevamine, a rubber tree protein with chitinase and lysozyme activities.
- Kluyveromyces lactis killer toxin alpha subunit, which acts as a chitinase.
- Flavobacterium and Streptomyces endo-beta-N-acetylglucosaminidases (EC 3.2.1.96).
- Mammalian di-N-acetylchitobiase which is involved in the degradation of asparagine-linked glycoproteins.
- Human cartilage glycoprotein Gp-39.
- Jack bean concanavalin B (conB), a protein that has lost its catalytic activity.

Site directed mutagenesis experiments [3] and crystallographic data [4,5] have shown that a conserved glutamate is involved in the catalytic mechanism and probably acts as a proton donor. This glutamate is at the extremity of the best conserved region in these proteins.

Consensus pattern[LIVMFY]-[DN]-G-[LIVMF]-[DN]-[LIVMF]-[DN]-x-E [E is the active site residue]

- [1] Flach J., Pilet P.-E., Jolles P. Experientia 48:701-716(1992).
- [2] Henrissat B. Biochem. J. 280:309-316(1991).
- [3] Watanabe T., Kohori K., Miyashita K., Fujii T., Sakai H., Uchida M., Tanaka H. J. Biol. Chem. 268:18567-18572(1993).
- [4] Perrakis A., Tews I., Dauter Z., Oppenheim A.B., Chet I., Wilson K.S., Vorgias C.E. Structure 2:1169-1180(1994).
- [5] van Scheltinga A.C.T., Kalk K.H., Beintema J.J., Dijkstra B.W. Structure 2:1181-1189(1994).

769. gag_p17. gag gene protein p17 (matrix protein).

The matrix protein forms an icosahedral shell associated with the inner membrane of the mature immunodeficiency virus. Number of members: 1598

[1] Medline: 95055757. Three-dimensional structure of the human immunodeficiency virus type 1 matrix protein. Massiah MA, Starich MR, Paschall C, Summers MF, Christensen AM, Sundquist WI; J Mol Biol 1994;244:198-223.

770. GDA1/CD39 family of nucleoside phosphatases signature

634

PROSITE cross-reference(s); GDA1_CD39_NTPASE

A number of nucleoside diphosphate and triphosphate hydrolases as well as some yet uncharacterized proteins have been found to belong to the same family [1, 2]. This family currently consist of:

- Yeast guanosine-diphosphatase (EC 3.6.1.42) (GDPase) (gene GDA1). GDA1 is a golgi integral membrane enzyme that catalyzes the hydrolysis of GDP to GMP.
- Potato apyrase (EC 3.6.1.5) (adenosine diphosphatase) (ADPase). Apyrase acts on both ATP and ADP to produce AMP.
- Mammalian vascular ATP-diphosphohydrolase (EC 3.6.1.5) (also known as lymphoid cell activation antigen CD39).
- Toxoplasma gondii nucleoside-triphosphatases (EC 3.6.1.15) (NTPase). NTPase hydrolyses various nucleoside triphosphates to produce the corresponding nucleoside mono- and diphosphates. This enzyme is secreted into the invaded host cell into the parasitophorous vacuole, a specialized compartment where the parasite intracellularly resides.
- Pea nucleoside-triphosphatases (EC 3.6.1.15) (NTPase).
- Caenorhabditis elegans hypothetical protein C33H5.14.
- Caenorhabditis elegans hypothetical protein R07E4.4.
- Yeast chromosome V hypothetical protein YER005w.

The above uncharacterized proteins all seem to be membrane-bound.

All these proteins share a number of conserved domains. The best conserved of these domains have been selected. It is located in the central section of the proteins.

Consensus pattern[LIVM]-x-G-x(2)-E-G-x-[FY]-x-[FW]-[LIVA]-[TAG]-x-N-[HY]

- [1] Handa M., Guidotti G. Biochem. Biophys. Res. Commun. 218:916-923(1996).
- [2] Vasconcelos E.G., Ferreira S.T., de Carvalho T.M.U., de Souza W., Kettlun A.M., Mancilla M., Valenzuela M.A., Verjovski-Almeida S. J. Biol. Chem. 271:22139-22145(1996).

771. GTP cyclohydrolase I signatures

PROSITE cross-reference(s); GTP_CYCLOHYDROL_1_1, GTP_CYCLOHYDROL_1_2

GTP cyclohydrolase I (EC 3.5.4.16) catalyzes the biosynthesis of formic acid and
 5 dihydroneopterin triphosphate from GTP. This reaction is the first step in the biosynthesis of
 tetrahydrofolate in prokaryotes, of tetrahydrobiopterin in vertebrates, and of pteridine-
 containing pigments in insects.

GTP cyclohydrolase I is a protein of from 190 to 250 amino acid residues. The comparison
 10 of the sequence of the enzyme from bacterial and eukaryotic sources shows that the
 structure of this enzyme has been extremely well conserved throughout evolution [1].

Two conserved regions were selected as signature patterns. The first contains a perfectly
 conserved tetrapeptide which is part of the GTP-binding pocket [2], the second region also
 15 contains conserved residues involved in GTP-binding.

Consensus pattern[DEN]-[LIVM](2)-x(2)-[KRNQ]-[DEN]-[LIVM]-x(3)-[ST]-x-C-E- H-H
 Consensus pattern[SA]-x-[RK]-x-Q-[LIVM]-Q-E-[RN]-[LI]-[TSN]

[1] Maier J., Witter K., Guetlich M., Ziegler I., Werner T., Ninnemann H. Biochem.
 20 Biophys. Res. Commun. 212:705-711(1995).

[2] Nar H., Huber R., Meining W., Schmid C., Weinkauff S., Bacher A. Structure 3:459-
 466(1995).

25 772. IlvC. Acetohydroxy acid isomeroreductase

Acetohydroxy acid isomeroreductase catalyses the conversion of acetohydroxy acids into
 dihydroxy valerates. This reaction is the second in the synthetic pathway of the essential
 branched side chain amino acids valine and isoleucine. Number of members: 29

[1] Medline: 97361822. The crystal structure of plant acetohydroxy acid isomeroreductase
 30 complexed with NADPH, two magnesium ions and a herbicidal transition state analog
 determined at 1.65 Å resolution. Biou V, Dumas R, Cohen-Addad C, Douce R, Job D, Pebay-
 Peyroula E; EMBO J 1997;16:3405-3415.

773. Prokaryotic membrane lipoprotein lipid attachment site

PROSITE cross-reference(s); PROKAR_LIPOPROTEIN

In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):

- Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp).
- Escherichia coli lipoprotein-28 (gene nlpA).
- Escherichia coli lipoprotein-34 (gene nlpB).
- Escherichia coli lipoprotein nlpC.
- Escherichia coli lipoprotein nlpD.
- Escherichia coli osmotically inducible lipoprotein B (gene osmB).
- Escherichia coli osmotically inducible lipoprotein E (gene osmE).
- Escherichia coli peptidoglycan-associated lipoprotein (gene pal).
- Escherichia coli rare lipoproteins A and B (genes rplA and rplB).
- Escherichia coli copper homeostasis protein cutF (or nlpE).
- Escherichia coli plasmids traT proteins.
- Escherichia coli Col plasmids lysis proteins.
- A number of Bacillus beta-lactamases.
- Bacillus subtilis periplasmic oligopeptide-binding protein (gene oppA).
- Borrelia burgdorferi outer surface proteins A and B (genes ospA and ospB).
- Borrelia hermsii variable major protein 21 (gene vmp21) and 7 (gene vmp7).
- Chlamydia trachomatis outer membrane protein 3 (gene omp3).
- Fibrobacter succinogenes endoglucanase cel-3.
- Haemophilus influenzae proteins Pal and Pcp.
- Klebsiella pullulunase (gene pulA).
- Klebsiella pullulunase secretion protein pulS.
- Mycoplasma hyorhina protein p37.
- Mycoplasma hyorhina variant surface antigens A, B, and C (genes vlpABC).
- Neisseria outer membrane protein H.8.
- Pseudomonas aeruginosa lipopeptide (gene lppL).

637

- *Pseudomonas solanacearum* endoglucanase egl.
- *Rhodopseudomonas viridis* reaction center cytochrome subunit (gene cytC).
- *Rickettsia* 17 Kd antigen.
- *Shigella flexneri* invasion plasmid proteins mxiJ and mxiM.
- 5 - *Streptococcus pneumoniae* oligopeptide transport protein A (gene amiA).
- *Treponema pallidum* 34 Kd antigen.
- *Treponema pallidum* membrane protein A (gene tmpA).
- *Vibrio harveyi* chitinase (gene chb).
- *Yersinia* virulence plasmid protein yscJ.

10

- Halocyanin from *Natrobacterium pharaonis* [4], a membrane associated copper-binding protein. This is the first archaeobacterial protein known to be modified in such a fashion).

From the precursor sequences of all these proteins, we derived a consensus pattern and a set of rules to identify this type of post-translational modification.

Consensus pattern{DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1) The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least one Lys or one Arg in the first seven positions of the sequence.

[1] Hayashi S., Wu H.C. J. Bioenerg. Biomembr. 22:451-471(1990).

[2]Klein P., Somorjai R.L., Lau P.C.K. Protein Eng. 2:15-20(1988).

[3]von Heijne G. Protein Eng. 2:531-534(1989).

25 [4]Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).

774. Aminoacyl-transfer RNA synthetases class-II signatures

PROSITE cross-reference(s); AA_TRNA_LIGASE_II_1; AA_TRNA_LIGASE_II_2

30 Aminoacyl-tRNA synthetases (EC 6.1.1.-) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are

generally two aminoacyl-tRNA synthetases for each different amino acid: one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure.

- 5 The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is different to the Rossmann fold observed for the class I synthetases [7].

10

Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. Signature patterns from two of these regions have been derived.

11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

Consensus pattern[FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE]

Consensus pattern[GSTALVF]-{DENQHRKP}-[GSTA]-[LIVMF]-[DE]-R-[LIVMF]-x-[LIVMSTAG]-[LIVMFY]

[1]Schimmel P. Annu. Rev. Biochem. 56:125-158(1987).

[2]Delarue M., Moras D. BioEssays 15:675-687(1993).

[3]Schimmel P. Trends Biochem. Sci. 16:1-3(1991).

[4]Nagel G.M., Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).

[5]Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).

[6]Cusack S. Biochimie 75:1077-1081(1993).

25 [7]Cusack S., Berthet-Colominas C., Haertlein M., Nassar N., Leberman R. Nature 347:249-255(1990).

[8]Leveque F., Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).

775. X. Trans-activation protein X

30 This protein is found in hepadnaviruses where it is indispensable for replication. Number of members: 91

776. Thymidylate synthase active site

Thymidylate synthase (EC 2.1.1.45) [1,2] catalyzes the reductive methylation of dUMP to dTMP with concomitant conversion of 5,10-methylenetetrahydrofolate to dihydrofolate. Thymidylate synthase plays an essential role in DNA synthesis and is an important target for certain chemotherapeutic drugs.

Thymidylate synthase is an enzyme of about 30 to 35 Kd in most species except in protozoan and plants where it exists as a bifunctional enzyme that includes a dihydrofolate reductase domain.

A cysteine residue is involved in the catalytic mechanism (it covalently binds the 5,6-dihydro-dUMP intermediate). The sequence around the active site of this enzyme is conserved from phages to vertebrates.

Consensus pattern R-x(2)-[LIVM]-x(3)-[FW]-[QN]-x(8,9)-[LV]-x-P-C-[HAVM]-x(3)-[QMT]-[FYW]-x-[LV] [C is the active site residue]

[1] Benkovic S.J. Annu. Rev. Biochem. 49:227-251(1980).

[2] Ross P., O'Gara F., Condon S. Appl. Environ. Microbiol. 56:2156-2163(1990).

777. Glycosyl hydrolases family 31 signatures

It has been shown [1,2,3,E1] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:

- Lysosomal alpha-glucosidase (EC 3.2.1.20) (acid maltase) is a vertebrate glycosidase active at low pH, which hydrolyzes alpha(1->4) and alpha(1->6) linkages in glycogen, maltose, and isomaltose.
- Alpha-glucosidase (EC 3.2.1.20) from the yeast *Candida tsukunbaensis*.
- Alpha-glucosidase (EC 3.2.1.20) (gene malA) from the archaebacteria *Sulfolobus solfataricus*.
- Intestinal sucrase-isomaltase (EC 3.2.1.48 / EC 3.2.1.10) is a vertebrate membrane-bound, multifunctional enzyme complex which hydrolyzes sucrose, maltose and isomaltose. The sucrase and isomaltase domains of the enzyme are homologous (41% of amino acid identity) and have most probably evolved by duplication.
- Glucoamylase 1 (EC 3.2.1.3) (glucan 1,4-alpha-glucosidase) from various fungal species.
- Yeast hypothetical protein YBR229c.
- Fission yeast hypothetical protein SpAC30D11.01c.

An aspartic acid has been implicated [4] in the catalytic activity of sucrase, isomaltase, and lysosomal alpha-glucosidase. The region around this active residue is highly conserved and can be used as a signature pattern. A second region, which contains two conserved cysteines, has been used as an additional signature pattern.

5

Consensus pattern [GF]-[LIVMF]-W-x-D-M-[NSA]-E [D is the active site residue]

Consensus pattern G-[AV]-D-[LIVMTA]-C-G-[FY]-x(3)-[ST]-x(3)-L-C-x-R-W-x(2)-[LV]-[GSA]-[SA]-F-x-P-F-x-R-[DN]

10

[1] Henrissat B. Biochem. J. 280:309-316(1991).

[2] Kinsella B.T., Hogan S., Larkin A., Cantwell B.A. Eur. J. Biochem. 202:657-664(1991).

[3] Naim H.Y., Niermann T., Kleinhans U., Hollenberg C.P., Strasser A.W.M. FEBS Lett. 294:109-112(1991).

[4] Hermans M.M.P., Kroos M.A., van Beeumen J., Oostra B.A., Reuser A.J.J. J. Biol. Chem. 266:13507-13512(1991).

15

778. Urease signatures

Urease (EC 3.5.1.5) is a nickel-binding enzyme that catalyzes the hydrolysis of urea to carbon dioxide and ammonia [1]. Historically, it was the first enzyme to be crystallized (in 1926). It is mainly found in plant seeds, microorganisms and invertebrates. In plants, urease is a hexamer of identical chains. In bacteria [2], it consists of either two or three different subunits (alpha, beta and gamma).

20

Urease binds two nickel ions per subunit; four histidine, an aspartate and a carbamated-lysine serve as ligands to these metals; an additional histidine is involved in the catalytic mechanism [3].

25

As signatures for this enzyme, a region was selected that contains two histidine that bind one of the nickel ions and the region of the active site histidine.

Consensus pattern T-[AY]-[GA]-[GAT]-[LIVM]-D-x-H-[LIVM]-H-x(3)-P [The two H's bind nickel]

30

Consensus pattern [LIVM](2)-[CT]-H-[HN]-L-x(3)-[LIVM]-x(2)-D-[LIVM]-x-F-A [H is the active site residue]

- [1] Takishima K., Suga T., Mamiya G. Eur. J. Biochem. 175:151-165(1988).
 [2] Mobley H.L.T., Husinger R.P. Microbiol. Rev. 53:85-108(1989).
 [3] Jabri E., Carr M.B., Hausinger R.P., Karplus P.A. Science 268:998-1004(1995).

5 779. Tyrosine specific protein phosphatases signature and profiles

Tyrosine specific protein phosphatases (EC 3.1.3.48) (PTPase) [1 to 5] are enzymes that catalyze the removal of a phosphate group attached to a tyrosine residue. These enzymes are very important in the control of cell growth, proliferation, differentiation and transformation. Multiple forms of PTPase have been characterized and can be classified into
 10 two categories: soluble PTPases and transmembrane receptor proteins that contain PTPase domain(s). The currently known PTPases are listed below:

Soluble PTPases.

- PTPN1 (PTP-1B).
- PTPN2 (T-cell PTPase; TC-PTP).
- PTPN3 (H1) and PTPN4 (MEG), enzymes that contain an N-terminal band 4.1- like domain (see <PDOC00566>) and could act at junctions between the membrane and cytoskeleton.
- PTPN5 (STEP).
- PTPN6 (PTP-1C; HCP; SHP) and PTPN11 (PTP-2C; SH-PTP3; Syp), enzymes which contain two copies of the SH2 domain at its N-terminal extremity. The Drosophila protein corkscrew (gene csw) also belongs to this subgroup.
- PTPN7 (LC-PTP; Hematopoietic protein-tyrosine phosphatase; HePTP).
- PTPN8 (70Z-PEP).
- 25 - PTPN9 (MEG2).
- PTPN12 (PTP-G1; PTP-P19).
- Yeast PTP1.
- Yeast PTP2 which may be involved in the ubiquitin-mediated protein degradation pathway.
- 30 - Fission yeast pyp1 and pyp2 which play a role in inhibiting the onset of mitosis.
- Fission yeast pyp3 which contributes to the dephosphorylation of cdc2.
- Yeast CDC14 which may be involved in chromosome segregation.
- Yersinia virulence plasmid PTPases (gene yopH).

- Autographa californica nuclear polyhedrosis virus 19 Kd PTPase.

Dual specificity PTPases.

- DUSP1 (PTPN10; MAP kinase phosphatase-1; MKP-1); which dephosphorylates MAP kinase on both Thr-183 and Tyr-185.
- DUSP2 (PAC-1), a nuclear enzyme that dephosphorylates MAP kinases ERK1 and ERK2 on both Thr and Tyr residues.
- DUSP3 (VHR).
- DUSP4 (HVH2).
- DUSP5 (HVH3).
- DUSP6 (Pyst1; MKP-3).
- DUSP7 (Pyst2; MKP-X).
- Yeast MSG5, a PTPase that dephosphorylates MAP kinase FUS3.
- Yeast YVH1.
- Vaccinia virus H1 PTPase; a dual specificity phosphatase.

Receptor PTPases.

Structurally, all known receptor PTPases, are made up of a variable length extracellular domain, followed by a transmembrane region and a C-terminal catalytic cytoplasmic domain. Some of the receptor PTPases contain fibronectin type III (FN-III) repeats, immunoglobulin-like domains, MAM domains or carbonic anhydrase-like domains in their extracellular region. The cytoplasmic region generally contains two copies of the PTPase domain. The first seems to have enzymatic activity, while the second is inactive but seems to affect substrate specificity of the first. In these domains, the catalytic cysteine is generally conserved but some other, presumably important, residues are not.

In the following table, the domain structure of known receptor PTPases is shown:

Extracellular	Intracellular				
-----	-----				
Ig FN-3	CAH	MAM	PTPase		
Leukocyte common antigen (LCA) (CD45)	0	2	0	0	2

		643				
	Leukocyte antigen related (LAR)	3	8	0	0	2
	Drosophila DLAR	3	9	0	0	2
	Drosophila DPTP	2	2	0	0	2
	PTP-alpha (LRP)	0	0	0	0	2
5	PTP-beta	0	16	0	0	1
	PTP-gamma	0	1	1	0	2
	PTP-delta	0	>7	0	0	2
	PTP-epsilon	0	0	0	0	2
	PTP-kappa	1	4	0	1	2
10	PTP-mu	1	4	0	1	2
	PTP-zeta	0	1	1	0	2

PTPase domains consist of about 300 amino acids. There are two conserved cysteines, the second one has been shown to be absolutely required for activity. Furthermore, a number of conserved residues in its immediate vicinity have also been shown to be important.

A signature pattern was derived for PTPase domains centered on the active site cysteine.

There are three profiles for PTPases, the first one spans the complete domain and is not specific to any subtype. The second profile is specific to dual-specificity PTPases and the third one to the PTP subfamily.

Consensus pattern [LIVMF]-H-C-x(2)-G-x(3)-[STC]-[STAGP]-x-[LIVMFY] [C is the active site residue]

Notethe M-phase inducer phosphatases (cdc25-type phosphatase) are tyrosine- protein phosphatases that are not structurally related to the above PTPases.

Notethis documentation entry is linked to both a signature pattern and to profiles. As profiles are much more sensitive than the pattern, you should use them if you have access to the necessary software tools to do so.

[1] Fischer E.H., Charbonneau H., Tonks N.K. Science 253:401-406(1991).

[2] Charbonneau H., Tonks N.K. Annu. Rev. Cell Biol. 8:463-493(1992).

[3] Trowbridge I.S. J. Biol. Chem. 266:23517-23520(1991).

[4] Tonks N.K., Charbonneau H. Trends Biochem. Sci. 14:497-500(1989).

[5] Hunter T. Cell 58:1013-1016(1989).

780. Connexins signatures

Gap junctions [1] are specialized regions of the plasma membrane which consist of closely packed pairs of transmembrane channels, the connexons, through which small molecules diffuse from a cell to a neighboring cell. Each connexon is composed of an hexamer of an integral membrane protein which is often referred to as connexin. In a given species there are a number of different, yet structurally related, tissue specific, forms of connexins. The types of connexins which are currently known are listed below.

- Connexin 56 (Cx56).
- Connexin 50 (Cx50) (lens fiber protein MP70).
- Connexin 46 (Cx46) (alpha-3).
- Connexin 45 (Cx45) (alpha-6).
- Connexin 43 (Cx43) (alpha-1).
- Connexin 40 (Cx40) (alpha-5).
- Connexin 38 (Cx38) (alpha-2).
- Connexin 37 (Cx37) (alpha-4).
- Connexin 33 (Cx33) (alpha-7).
- Connexin 32 (Cx32) (beta-1).
- Connexin 31.1 (Cx31.1) (beta-4).
- Connexin 31 (Cx31) (beta-3).
- Connexin 30.3 (Cx30.3) (beta-5).
- Connexin 26 (Cx26) (beta-2).

Structurally the connexins consist of a short cytoplasmic N-terminal domain, followed by four transmembrane segments that delimit two extracellular and one cytoplasmic loops; the C-terminal domain is cytoplasmic and its length is variable (from 20 residues in Cx26 to 260 residues in Cx56). The schematic representation of this structure is shown below.

```

NH2-***      ***      *****-COOH
      **      **      **      **
      **      **      **      **  Cytoplasmic
      ---**---**---**---**-----
      **      **      **      **  Membrane
      **      **      **      **

```

645

```

-----**-----**-----**-----**-----
**  **  **  **  Extracellular
** **  ** **
**      **

```

5 The sequences of the two extracellular loops are well conserved. In both loops there are three conserved cysteines which are involved in disulfide bonds. A signature patterns from each of these two loop regions has been built.

10 Consensus pattern C-[DN]-T-x-Q-P-G-C-x(2)-V-C-[FY]-D [The three C's are involved in disulfide bonds] Consensus pattern C-x(3,4)-P-C-x(3)-[LIVM]-[DEN]-C-[FY]-[LIVM]-[SA]-[KR]-P [The three C's are involved in disulfide bonds]

[1] Goodenough D.A., Goliger J.A., Paul D.L. Annu. Rev. Biochem. 65:475-502(1996).

15 781. Gram-positive cocci surface proteins 'anchoring' hexapeptide

Surface proteins from Gram-positive cocci contains a conserved hexapeptide located a few residues downstream of a hydrophobic C-terminal membrane anchor region which is followed by a cluster of basic amino acids [1]. This structure is represented in the following schematic representation:

```

+-----+-----+-----+-----+-----+-----+
| Variable length extracellular domain  [H] Anchor [B]
+-----+-----+-----+-----+-----+-----+

```

'H': conserved hexapeptide.

25 'B': cluster of basic residues.

It has been proposed that this hexapeptide sequence is responsible for a post-translational modification necessary for the proper anchoring of the proteins which bear it, to the cell wall.

Proteins known to contain such hexapeptide are listed below:

- 30
- Aggregation substance from streptococcus faecalis (asa1).
 - C5a peptidase from Streptococcus pyogenes (scpA).
 - C protein alpha-antigen from Streptococcus agalactiae (bca).
 - Cell surface antigen I/II (PAC) from Streptococcus mutans.

646

- Dextranase from *Streptococcus downei* (dex).
- Fibronectin-binding protein from *Staphylococcus aureus* (fnbA).
- Fimbrial subunits from *Actinomyces naeslundii* and *viscosus*.
- IgA binding protein from *Streptococcus pyogenes* (arp4).
- 5 - IgA binding protein (B antigen) from *Streptococcus agalactiae* (bag).
- IgG binding proteins from *Streptococci* and *Staphylococcus aureus*.
- Internalin A from *Listeria monocytogenes* (inlA).
- M proteins from streptococci.
- Muramidase-released protein from *Streptococcus suis* (mrp).
- 10 - Nisin leader peptide processing protease from *Lactococcus lactis* (nisP).
- Protein A from *Staphylococcus aureus*.
- Trypsin-resistant surface T protein from streptococci.
- Wall-associated protein from *Streptococcus mutans* (wapA).
- Wall-associated serine proteinases from *Lactococcus lactis*.

Consensus pattern L-P-x-T-G-[STGAVDE]

[1] Schneewind O., Jones K.F., Fischetti V.A. J. Bacteriol. 172:3310-3317(1990).

782. Gamma-glutamyltranspeptidase signature

Gamma-glutamyltranspeptidase (EC 2.3.2.2) (GGT) [1] catalyzes the transfer of the gamma-glutamyl moiety of glutathione to an acceptor that may be an amino acid, a peptide or water (forming glutamate). GGT plays a key role in the gamma-glutamyl cycle, a pathway for the synthesis and degradation of glutathione. In prokaryotes and eukaryotes, it is an enzyme that consists of two polypeptide chains, a heavy and a light subunit, processed from a single chain precursor. The active site of GGT is known to be located in the light subunit.

The sequences of mammalian and bacterial GGT show a number of regions of high similarity [2]. *Pseudomonas cephalosporin acylases* (EC 3.5.1.-) that convert 7-beta-(4-carboxybutanamido)-cephalosporanic acid (GL-7ACA) into 7-aminocephalosporanic acid (7ACA) and glutaric acid are evolutionary related to GGT and also show some GGT activity [3]. Like GGT, these GL-7ACA acylases, are also composed of two subunits.

One of the conserved regions correspond to the N-terminal extremity of the mature light chains of these enzymes. This region has been used as a signature pattern.

Consensus pattern T-[STA]-H-x-[ST]-[LIVMA]-x(4)-G-[SN]-x-V-[STA]-x-T-x-T-[LIVM]-
[NE]-x(1,2)-[FY]-G

- 5 [1] Tate S.S., Meister A. Meth. Enzymol. 113:400-419(1985).
[2] Suzuki H., Kumagai H., Echigo T., Tochikura T. J. Bacteriol. 171:5169-5172(1989).
[3] Ishiye M., Niwa M. Biochim. Biophys. Acta 1132:233-239(1992).

783. Ferrochelatase signature

10 Ferrochelatase (EC 4.99.1.1) (protoheme ferro-lyase) [1,2] catalyzes the last step in heme biosynthesis: the chelation of a ferrous ion to proto-porphyrin IX, to form protoheme.

In eukaryotes, ferrochelatase is a mitochondrial protein bound to the inner membrane, whose active site faces the mitochondrial matrix. The mature form of eukaryotic ferrochelatase is composed of about 360 amino acids. In bacteria, ferrochelatase (gene hemH) 15 [3] is a protein of from 310 to 380 amino acids.

The human autosomal dominant disease protoporphyria is due to the reduced activity of ferrochelatase.

The signature pattern for this enzyme is based on a conserved region which contains a histidine residue which could be involved in binding iron.

20 Consensus pattern [LIVMF](2)-x-[ST]-x-H-[GS]-[LIVM]-P-x(4,5)-[DENQKR]-x-G-[DP]-x(1,2)-Y

- [1] Labbe-Bois R. J. Biol. Chem. 265:7278-7283(1990).
25 [2] Brenner D.A., Frasier F. Proc. Natl. Acad. Sci. U.S.A. 88:849-853(1991).
[3] Miyamoto K., Nakahigashi K., Nishimura K., Inokuchi H. J. Mol. Biol. 219:393-398(1991).

784. Cellulose-binding domain, bacterial type

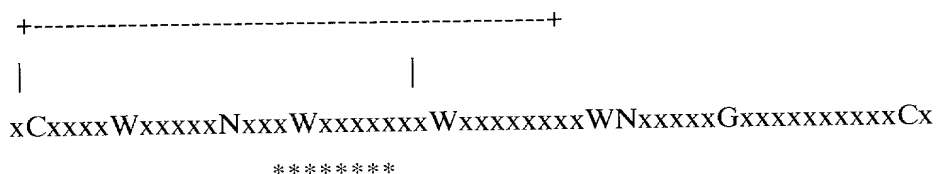
30 The microbial degradation of cellulose and xylans requires several types of enzyme such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1].

Structurally, cellulases and xylanases generally consist of a catalytic domain joined to a cellulose-binding domain (CBD) by a short linker sequence rich in proline and/or hydroxy-amino acids.

The CBD of a number of bacterial cellulases has been shown to consist of about 105 amino acid residues [2]. Enzymes known to contain such a domain are:

- Endoglucanase (gene end1) from *Butyrivibrio fibrisolvens*.
- Endoglucanases A (gene cenA) and B (cenB) from *Cellulomonas fimi*.
- Exoglucanases A (gene cbhA) and B (cbhB) from *Cellulomonas fimi*.
- Endoglucanase E-2 (gene celB) from *Thermomonospora fusca*.
- Endoglucanase A (gene celA) from *Microbispora bispora*.
- Endoglucanases A (gene celA), B (celB) and C (celC) from *Pseudomonas fluorescens*.
- Endoglucanase A (gene celA) from *Streptomyces lividans*.
- Exocellobiohydrolase (gene cex) from *Cellulomonas fimi*.
- Xylanases A (gene xynA) and B (xynB) from *Pseudomonas fluorescens*.
- Arabinofuranosidase C (EC 3.2.1.55) (xylanase C) (gene xynC) from *Pseudomonas fluorescens*.
- Chitinase 63 (EC 3.2.1.14) from *Streptomyces plicatus*.
- Chitinase C from *Streptomyces lividans*.

The CBD domain is found either at the N-terminal or at the C-terminal extremity of these enzymes. As it is shown in the following schematic representation, there are two conserved cysteines in this CBD domain - one at each extremity of the domain - which have been shown [3] to be involved in a disulfide bond. There are also four conserved tryptophan residues which could be involved in the interaction of the CBD with polysaccharides.



'C': conserved cysteine involved in a disulfide bond. '*': position of the pattern.

Consensus pattern W-N-[STAGR]-[STDN]-[LIVM]-x(2)-[GST]-x-[GST]-x(2)-[LIVMFT]-[GA]

[1] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).

[2] Meinke A., Gilkes N.R., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Protein Seq. Data Anal. 4:349-353(1991).

[3] Gilkes N.R., Claeysens M., Aebersold R., Henrissat B., Meinke A., Morrison H.D., Kilburn D.G., Warren R.A.J., Miller R.C. Jr. Eur. J. Biochem. 202:367-377(1991).

785. Amidases signature

It has been shown [1,2,3] that several enzymes from various prokaryotic and eukaryotic organisms which are involved in the hydrolysis of amides (amidases) are evolutionary related. These enzymes are listed below.

- Indoleacetamide hydrolase (EC 3.5.1.-), a bacterial plasmid-encoded enzyme that catalyzes the hydrolysis of indole-3-acetamide (IAM) into indole-3-acetate (IAA), the second step in the biosynthesis of auxins from tryptophan.

- Acetamidase from *Emericella nidulans* (gene *amdS*), an enzyme which allows acetamide to be used as a sole carbon or nitrogen source.

- Amidase (EC 3.5.1.4) from *Rhodococcus* sp. N-774 and *Brevibacterium* sp. R312 (gene *amdA*). This enzyme hydrolyzes propionamides efficiently, and also at a lower efficiency, acetamide, acrylamide and indoleacetamide.

- Amidase (EC 3.5.1.4) from *Pseudomonas chlororaphis*.

- 6-aminohexanoate-cyclic-dimer hydrolase (EC 3.5.2.12) (nylon oligomers degrading enzyme E1) (gene *nylA*), a bacterial plasmid encoded enzyme which catalyzes the first step in the degradation of 6-aminohexanoic acid cyclic dimer, a by-product of nylon manufacture [4].

- Glutamyl-tRNA(Gln) amidotransferase subunit A [5].

- Mammalian fatty acid amide hydrolase (gene *FAAH*) [6].

- A putative amidase from yeast (gene *AMD2*).

- *Mycobacterium tuberculosis* putative amidases *amiA2*, *amiB2*, *amiC* and *amiD*.

All these enzymes contain in their central section a highly conserved region rich in glycine, serine, and alanine residues. This region has been used as a signature pattern.

Consensus pattern: G-[GA]-S-[GS]-[GS]-G-x-[GSA]-[GSAVY]-x-[LIVM]-[GSA]-x(6)-
[GSAT]-x-[GA]-x-[DE]-x-[GA]-x-S-[LIVM]-R-x-P-[GSAC]

[1] Mayaux J.-F., Cerbelaud E., Soubrier F., Faucher D., Petre D. J. Bacteriol. 172:6764-
5 6773(1990).

[2] Hashimoto Y., Nishiyama M., Ikehata O., Horinouchi S., Beppu T. Biochim. Biophys.
Acta 1088:225-233(1991).

[3] Chang T.-H., Abelson J. Nucleic Acids Res. 18:7180-7180(1990).

[4] Tsuchiya K., Fukuyama S., Kanzaki N., Kanagawa K., Negoro S., Okada H. J. Bacteriol.
10 171:3187-3191(1989).

[5] Curnow A.W., Hong K.W., Yuan R., Kim S.I., Martins O., Winkler W., Henkin T.M.,
Soll D. Proc. Natl. Acad. Sci. U.S.A. 94:11819-11826(1997).

[6] Cravatt B.F., Giang D.K., Mayfield S.P., Boger D.L., Lerner R.A., Gilula N.B. Nature
384:83-87(1996).

786. Glycosyl hydrolases family 10 active site

The microbial degradation of cellulose and xylans requires several types of enzymes
such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or
xylanases (EC 3.2.1.8) [1,2]. Fungi and bacteria produces a spectrum of cellulolytic enzymes
(cellulases) and xylanases which, on the basis of sequence similarities, can be classified into
families. One of these families is known as the cellulase family F [3] or as the glycosyl
hydrolases family 10 [4,E1]. The enzymes which are currently known to belong to this
family are listed below.

- *Aspergillus awamori* xylanase A (xynA).

- *Bacillus* sp. strain 125 xylanase (xynA).

- *Bacillus stearothermophilus* xylanase.

- *Butyrivibrio fibrisolvens* xylanases A (xynA) and B (xynB).

- *Caldocellum saccharolyticum* bifunctional endoglucanase/exoglucanase (celB). This
protein consists of two domains; it is the N-terminal domain, which has exoglucanase
activity, which belongs to this family.

- *Caldocellum saccharolyticum* xylanase A (xynA).

- *Caldocellum saccharolyticum* ORF4. This hypothetical protein is encoded in the xynABC
operon and is probably a xylanase.

651

- *Cellulomonas fimi* exoglucanase/xylanase (cex).
- *Clostridium stercorarium* thermostable cellobiohydrolase.
- *Clostridium thermocellum* xylanases Y (xynY) and Z (xynZ).
- *Cryptococcus albidus* xylanase.
- 5 - *Penicillium chrysogenum* xylanase (gene xylP).
- *Pseudomonas fluorescens* xylanases A (xynA) and B (xynB).
- *Ruminococcus flavefaciens* bifunctional xylanase XYLA (xynA). This protein consists of three domains: a N-terminal xylanase catalytic domain that belongs to family 11 of glycosyl hydrolases; a central domain composed of short repeats of Gln, Asn and Trp, and a C-terminal
- 10 xylanase catalytic domain that belongs to family 10 of glycosyl hydrolases.
- *Streptomyces lividans* xylanase A (xlnA).
- *Thermoanaerobacter saccharolyticus* endoxylanase A (xynA).
- *Thermoascus aurantiacus* xylanase.
- Thermophilic bacterium Rt8.B4 xylanase (xynA).

One of the conserved regions in these enzymes is centered on a conserved glutamic acid residue which has been shown [5], in the exoglucanase from *Cellulomonas fimi*, to be directly involved in glycosidic bond cleavage by acting as a nucleophile. This region has been used as a signature pattern.

Consensus pattern[GTA]-x(2)-[LIVN]-x-[IVMF]-[ST]-E-[LIY]-[DN]-[LIVMF] [E is the active site residue]

[1] Beguin P. Annu. Rev. Microbiol. 44:219-248(1990).

25 [2] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol. Rev. 55:303-315(1991).

[3] Henrissat B., Claeyssens M., Tomme P., Lemesle L., Mornon J.-P. Gene 81:83-95(1989).

[4] Henrissat B. Biochem. J. 280:309-316(1991).

30 [5] Tull D., Withers S.G., Gilkes N.R., Kilburn D.G., Warren R.A.J., Aebersold R. J. Biol. Chem. 266:15621-15625(1991).

787. Fructose-bisphosphate aldolase class-II signatures

Fructose-bisphosphate aldolase (EC 4.1.2.13) [1,2] is a glycolytic enzyme that catalyzes the reversible aldol cleavage or condensation of fructose-1,6- bisphosphate into dihydroxyacetone-phosphate and glyceraldehyde 3-phosphate. There are two classes of fructose-bisphosphate aldolases with different catalytic mechanisms. Class-II aldolases [2],
 5 mainly found in prokaryotes and fungi, are homodimeric enzymes which require a divalent metal ion – generally zinc - for their activity.

This family also includes the following proteins:

- Escherichia coli galactitol operon protein gatY which catalyzes the transformation of
 10 tagatose 1,6-bisphosphate into glycerone phosphate and D- glyceraldehyde 3-phosphate.
- Escherichia coli N-acetyl galactosamine operon protein agaY which catalyzes the same reaction as that of gatY.

As signature patterns for this class of enzyme, two conserved regions were selected. The first pattern is located in the first half of the sequence and contains two histidine residues that have been shown [4] to be involved in binding a zinc ion. The second is located in the C-terminal section and contains clustered acidic residues and glycines.

Consensus pattern[FYVMT]-x(1,3)-[LIVMH]-[APN]-[LIVM]-x(1,2)-[LIVM]-H-x-D-H-
 20 [GACH] [The two H's are zinc ligands]

Consensus pattern[LIVM]-E-x-E-[LIVM]-G-x(2)-[GM]-[GSTA]-x-E

[1] Perham R.N. Biochem. Soc. Trans. 18:185-187(1990).

[2] Marsh J.J., Lebherz H.G. Trends Biochem. Sci. 17:110-113(1992).

25 [3] von der Osten C.H., Barbas C.F. III, Wong C.-H., Sinskey A.J. Mol. Microbiol. 3:1625-1637(1989).

[4] Berry A., Marshall K.E. FEBS Lett. 318:11-16(1993).

788. Prolyl oligopeptidase family serine active site

30 The prolyl oligopeptidase family [1,2,3] consist of a number of evolutionary related peptidases whose catalytic activity seems to be provided by a charge relay system similar to that of the trypsin family of serine proteases, but which evolved by independent convergent evolution. The known members of this family are listed below.

- Prolyl endopeptidase (EC 3.4.21.26) (PE) (also called post-proline cleaving enzyme). PE is an enzyme that cleaves peptide bonds on the C-terminal side of prolyl residues. The sequence of PE has been obtained from a mammalian species (pig) and from bacteria (*Flavobacterium meningosepticum* and *Aeromonas hydrophila*); there is a high degree of sequence

conservation between these sequences.

- *Escherichia coli* protease II (EC 3.4.21.83) (oligopeptidase B) (gene prtB) which cleaves peptide bonds on the C-terminal side of lysyl and arginyl residues.

- Dipeptidyl peptidase IV (EC 3.4.14.5) (DPP IV). DPP IV is an enzyme that removes N-terminal dipeptides sequentially from polypeptides having unsubstituted N-termini provided that the penultimate residue is proline.

- Yeast vacuolar dipeptidyl aminopeptidase A (DPAP A) (gene: STE13) which is responsible for the proteolytic maturation of the alpha-factor precursor.

- Yeast vacuolar dipeptidyl aminopeptidase B (DPAP B) (gene: DAP2).

- Acylamino-acid-releasing enzyme (EC 3.4.19.1) (acyl-peptide hydrolase). This enzyme catalyzes the hydrolysis of the amino-terminal peptide bond of an N-acetylated protein to generate a N-acetylated amino acid and a protein with a free amino-terminus.

A conserved serine residue has experimentally been shown (in *E. coli* protease II as well as in pig and bacterial PE) to be necessary for the catalytic mechanism. This serine, which is part of the catalytic triad (Ser, His, Asp), is generally located about 150 residues away from the C-terminal extremity of these enzymes (which are all proteins that contains about 700 to 800 amino acids).

Consensus pattern D-x(3)-A-x(3)-[LIVMFYW]-x(14)-G-x-S-x-G-G-[LIVMFYW](2) [S is the active site residue]

Note these proteins belong to families S9A/S9B/S9C in the classification of peptidases [4,E1].

[1] Rawlings N.D., Polgar L., Barrett A.J. *Biochem. J.* 279:907-911(1991).

[2] Barrett A.J., Rawlings N.D. *Biol. Chem. Hoppe-Seyler* 373:353-360(1992).

[3] Polgar L., Szabo E.
Biol. Chem. Hoppe-Seyler 373:361-366(1992).

[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

789. Formate--tetrahydrofolate ligase signatures

Formate--tetrahydrofolate ligase (EC 6.3.4.3) (formyltetrahydrofolate synthetase) (FTHFS) is one of the enzymes participating in the transfer of one-carbon units, an essential element of various biosynthetic pathways. In many of these processes the transfers of one-carbon units are mediated by the coenzyme tetrahydrofolate (THF). Various reactions generate one-carbon derivatives of THF which can be interconverted between different oxidation states by FTHFS, methylenetetrahydrofolate dehydrogenase (EC 1.5.1.5) and methenyltetrahydrofolate cyclohydrolase (EC 3.5.4.9).

In eukaryotes the FTHFS activity is expressed by a multifunctional enzyme, C-1-tetrahydrofolate synthase (C1-THF synthase), which also catalyzes the dehydrogenase and cyclohydrolase activities. Two forms of C1-THF synthases are known [1], one is located in the mitochondrial matrix, while the second one is cytoplasmic. In both forms the FTHFS domain consist of about 600 amino acid residues and is located in the C-terminal section of C1-THF synthase. In prokaryotes FTHFS activity is expressed by a monofunctional homotetrameric enzyme of about 560 amino acid residues [2].

The sequence of FTHFS is highly conserved in all forms of the enzyme. As signature patterns, two regions that are almost perfectly conserved were selected. The first one is a glycine-rich segment located in the N-terminal part of FTHFS and which could be part of an ATP-binding domain [2]. The second pattern is located in the central section of FTHFS.

Consensus pattern G-[LIVM]-K-G-G-A-A-G-G-G-Y

Consensus pattern V-A-T-[IV]-R-A-L-K-x-[HN]-G-G

[1] Shannon K.W., Rabinowitz J.C. J. Biol. Chem. 263:7717-7725(1988).

[2] Lovell C.R., Przybyla A., Ljungdahl L.G. Biochemistry 29:5687-5694(1990).

790. Transthyretin signatures

Transthyretin (prealbumin) [1] is a thyroid hormone-binding protein that seems to transport thyroxine (T4) from the bloodstream to the brain. It is a protein of about 130 amino acids that assembles as a homotetramer and forms an internal channel that binds thyroxine.

Transthyretin is mainly synthesized in the brain choroid plexus. In humans, variants of the protein are associated with distinct forms of amyloidosis.

The sequence of transthyretin is highly conserved in vertebrates. A number of uncharacterized proteins also belong to this family:

- Escherichia coli hypothetical protein yedX.
- Bacillus subtilis hypothetical protein yunM.
- Caenorhabditis elegans hypothetical protein R09H10.3.
- Caenorhabditis elegans hypothetical protein ZK697.8.

Two regions were selected as signature patterns. The first located in the N-terminal extremity starts with a lysine known to be involved in binding T4. The second pattern is located in the C-terminal extremity.

Consensus pattern[KH]-[IV]-L-[DN]-x(3)-G-x-P-A-x(2)-[IV]-x-[IV] [The K binds thyroxine]

Consensus patternY-[TH]-[IV]-[AP]-x(2)-L-S-[PQ]-[FYW]-[GS]-[FY]-[QS]

[1] Schreiber G., Richardson S.J. Comp. Biochem. Physiol. 116B:137-160(1997).

791. Dihydropteroate synthase signatures

All organisms require reduced folate cofactors for the synthesis of a variety of metabolites. Most microorganisms must synthesize folate de novo because they lack the active transport system of higher vertebrate cells which allows these organisms to use dietary folates. Enzymes that are involved in the biosynthesis of folates are therefore the target of a variety of antimicrobial agents such as trimethoprim or sulfonamides.

Dihydropteroate synthase (EC 2.5.1.15) (DHPS) catalyzes the condensation of 6-hydroxymethyl-7,8-dihydropteridine pyrophosphate to para-aminobenzoic acid to form 7,8-dihydropteroate. This is the second step in the three steps pathway leading from 6-hydroxymethyl-7,8-dihydropterin to 7,8-dihydrofolate. DHPS is the target of sulfonamides which are substrates analog that compete with para-aminobenzoic acid.

Bacterial DHPS (gene sul or folP) [1] is a protein of about 275 to 315 amino acid residues which is either chromosomally encoded or found on various antibiotic resistance plasmids. In the lower eukaryote Pneumocystis carinii, DHPS is the C-terminal domain of a multifunctional folate synthesis enzyme (gene fas) [2].

Two signature patterns for DHPS were developed, the first signature is located in the N-terminal section of these enzymes, while the second signature is located in the central section.

- 5 Consensus pattern[LIVM]-x-[AG]-[LIVMF](2)-N-x-T-x-D-S-F-x-D-x-[SG]
Consensus pattern[GE]-[SA]-x-[LIVM](2)-D-[LIVM]-G-[GP]-x(2)-[STA]-x-P

[1] Slock J., Stahly D.P., Han C.-Y., Six E.W., Crawford I.P. J. Bacteriol. 172:7211-7226(1990).

- 10 [2] Volpes F., Dyer M., Scaife J.G., Darby G., Stammers D.K., Delves C.J. Gene 112:213-218(1992).

792. Phosphatidylinositol 3- and 4-kinases signatures

Phosphatidylinositol 3-kinase (PI3-kinase) (EC 2.7.1.137) [1] is an enzyme that phosphorylates phosphoinositides on the 3-hydroxyl group of the inositol ring. The exact function of the three products of PI3-kinase - PI-3-P, PI-3,4-P(2) and PI-3,4,5-P(3) - is not yet known, although it is proposed that they function as second messengers in cell signalling. Currently, three forms of PI3-kinase are known:

- The mammalian enzyme which is a heterodimer of a 110 Kd catalytic chain (p110) and an 85 Kd subunit (p85) which allows it to bind to activated tyrosine protein kinases. There are at least two different types of p100 subunits (alpha and beta).

- Yeast TOR1/DRR1 and TOR2/DRR2 [2], PI3-kinases required for cell cycle activation. Both are proteins of about 280 Kd.

- Yeast VPS34 [3], a PI3-kinase involved in vacuolar sorting and segregation. VPS34 is a protein of about 100 Kd.

- Arabidopsis thaliana and soybean VPS34 homologs.

Phosphatidylinositol 4-kinase (PI4-kinase) (EC 2.7.1.67) [4] is an enzyme that acts on phosphatidylinositol (PI) in the first committed step in the production of the second messenger inositol-1,4,5,-trisphosphate. Currently the following forms of PI4-kinases are known:

- Human PI4-kinase alpha.

- Yeast PIK1, a nuclear protein of 120 Kd.

- Yeast STT4, a protein of 214 Kd.

The PI3- and PI4-kinases share a well conserved domain at their C-terminal section; this domain seems to be distantly related to the catalytic domain of protein kinases [2]. Two signature patterns were developed from the best conserved parts of this domain.

Four additional proteins belong to this family:

- Mammalian FKBP-rapamycin associated protein (FRAP) [5], which acts as the target for the cell-cycle arrest and immunosuppressive effects of the FKBP12-rapamycin complex.
- Yeast protein ESR1 [6] which is required for cell growth, DNA repair and meiotic recombination.
- Yeast protein TEL1 which is involved in controlling telomere length.
- Yeast hypothetical protein YHR099w, a distantly related member of this family.
- Fission yeast hypothetical protein SpAC22E12.16C.

Consensus pattern[LIVMFAC]-K-x(1,3)-[DEA]-[DE]-[LIVMC]-R-Q-[DE]-x(4)-Q

Consensus pattern[GS]-x-[AV]-x(3)-[LIVM]-x(2)-[FYH]-[LIVM](2)-x-[LIVMF]-x-D-R-H-x(2)-N

[1] Hiles I.D., Otsu M., Volinia S., Fry M.J., Gout I., Dhand R., Panayotou G., Ruiz-Larrea F., Thompson A., Totty N.F., Hsuan J.J., Courtneidge S.A., Parker P.J., Waterfield M.D. Cell 70:419-429(1992).

[2] Kunz J., Henriquez R., Schneider U., Deuter-Reinhard M., Movva N., Hall M.N. Cell 73:585-596(1993).

[3] Schu P.V., Takegawa K., Fry M.J., Stack J.H., Waterfield M.D., Emr S.D. Science 260:88-91(1993).

[4] Garcia-Bustos J.F., Marini F., Stevenson I., Frei C., Hall M.N. EMBO J. 13:2352-2361(1994).

[5] Brown E.J., Albers M.W., Shin T.B., Ichikawa K., Keith C.T., Lane W.S., Schreiber S.L. Nature 369:756-758(1994).

[6] Kato R., Ogawa H. Nucleic Acids Res. 22:3104-3112(1994).

793. FAD-dependent glycerol-3-phosphate dehydrogenase signatures

FAD-dependent glycerol-3-phosphate dehydrogenase (EC 1.1.99.5) (GPD) catalyzes the conversion of glycerol-3-phosphate into dihydroxyacetone phosphate. In bacteria [1] it is associated with the utilization of glycerol coupled to respiration. In *Escherichia coli*, two isozymes are known: one expressed under anaerobic conditions (gene *glpA*) and one in aerobic conditions (gene *glpD*). In eukaryotes, a mitochondrial form of GPD participates in the glycerol phosphate shuttle in conjunction with an NAD-dependent cytoplasmic GPD (EC 1.1.1.8) [2,3].

These enzymes are proteins of about 60 to 70 Kd which contain a probable FAD-binding domain in their N-terminal extremity. The mammalian enzyme differs from the bacterial or yeast proteins by having an EF-hand calcium-binding region (See <PDOC00018>) in its C-terminal extremity.

Two signature patterns were developed. One based on the first half of the FAD-binding domain and one which corresponds to a conserved region in the central part of these enzymes.

Consensus pattern[IV]-G-G-G-x(2)-G-[STACV]-G-x-A-x-D-x(3)-R-G

Consensus patternG-G-K-x(2)-[GSTE]-Y-R-x(2)-A

[1] Austin D., Larson T.J. J. Bacteriol. 173:101-107(1991).

[2] Roennow B., Kielland-Brandt M.C. Yeast 9:1121-1130(1993).

[3] Brown L.J., McDonald M.J., Lehn D.A., Moran S.M. J. Biol. Chem. 269:14363-14366(1994).

794. NOL1/NOP2/sun family signature

The following proteins seems to be evolutionary related:

- Mammalian proliferating-cell nucleolar antigen p120 (gene NOL1) which may play a role in the regulation of the cell cycle and the increased nucleolar activity that is associated with the cell proliferation.

- Yeast nucleolar protein NOP2 (or YNA1) which could be involved in nucleolar function during the onset of growth, and in the maintenance of nucleolar structure.

- Yeast hypothetical protein YBL024w.

- Bacterial protein sun (also known as *fmU*).

- *Escherichia coli* hypothetical protein *yebU*.

659

- Mycobacterium tuberculosis hypothetical protein MtCY21B4.24.
- Methanococcus jannaschii hypothetical protein MJ0026.

NOL1 is a protein of 855 residues, NOP2 consists of 618 residues, YBL024w of 684, sun is a protein of about 430 to 450 residues and MJ026 has 274 residues. They share a conserved central domain which contains some highly conserved regions. One of these regions was selected as a signature pattern.

Consensus pattern[FV]-D-[KRA]-[LIVMA]-L-x-D-[AV]-P-C-[ST]-[GA]

795. moaA / nifB / pqqE family signature

A number of proteins involved in the biosynthesis of metallo cofactors have been shown [1,2] to be evolutionary related. These proteins are:

- Bacterial and archebacterial protein moaA, which is involved in the biosynthesis of the molybdenum cofactor (molybdopterin; MPT).
- Arabidopsis thaliana cnx2, a protein involved in molybdopterin biosynthesis and which is highly similar to moaA.
- Bacillus subtilis narA, which seems to be the moaA ortholog in that bacteria.
- Bacterial protein nifB (or fixZ) which is involved in the biosynthesis of the nitrogenase iron-molybdenum cofactor.
- Bacterial protein pqqE which is involved in the biosynthesis of the cofactor pyrrolo-quinoline-quinone (PQQ).
- Pyrococcus furiosus cmo, a protein involved in the synthesis of a molybdopterin-based tungsten cofactor.
- Caenorhabditis elegans hypothetical protein F49E2.1.

All these proteins share, in their N-terminal region, a conserved domain that contains three cysteines. In moaA, these cysteines have been shown [1] to be important for the biological activity. They could be involved in the binding of an iron-sulfur cluster.

Consensus pattern[LIV]-x(3)-C-[NP]-[LIVMF]-[QRS]-C-x-[FYM]-C [The three C's are putative Fe-S ligands]

- [1] Menendez C., Igloi G., Henninger H., Brandsch R. Arch. Microbiol. 164:142-151(1995).
[2] Hoff T., Schnorr K.M., Meyer C., Caboche M. J. Biol. Chem. 270:6100-6107(1995).

796. Forkhead-associated (FHA) domain profile

5 The forkhead-associated (FHA) domain [1,E1] is a putative nuclear signalling domain found in a variety of otherwise unrelated proteins. The FHA domain comprise approximately 55 to 75 amino acids and contains three highly conserved blocks separated by divergent spacer regions. Currently it has been found in the following proteins:

- Four transcription factors that also contain a forkhead (FH) domain: mouse myocyte
10 nuclear factor 1 (MNF1), yeast transcription factor FHL1, which probably controls pre-mRNA processing, and yeast FKH1 and FKH2. In those protein the FHA domain is located N-terminal of the DNA-binding FH domain.

- Kinase-associated protein phosphatase (KAPP) from Arabidopsis thaliana, a protein which
15 specifically interacts with the receptor-type Ser/Thr-kinase RLK5. In KAPP, the FHA domain maps to a region that interacts with the receptor-type protein kinase RLK5 only if the kinase is phosphorylated on serine residues [2].

- Two protein kinases from yeast that are involved in mediating the nuclear response to DNA
20 damage: DUN1 and SPK1/SAD1 [3]. The latter is the only known protein containing two copies of the FHA domain.

- Protein kinase cds1 from fission yeast contains a FHA domain and might be the ortholog of
SPK1.

- Protein kinase MEK1 from yeast, which is involved in meiotic recombination.

- Human nuclear antigen Ki67 which is expressed only in proliferating cells.

- Yeast hypothetical protein YHR115c, which contains a RING-finger C-terminal of the
25 FHA domain.

- Yeast hypothetical proteins L8083.1 and 9346.10, which contain an extensive coiled-coil
region C-terminal of the FHA domain.

- Caenorhabditis elegans hypothetical protein ZK632.2.

- Caenorhabditis elegans hypothetical protein C01G6.5.

30 - FraH from the prokaryote Anabaena, which contains a zinc-finger motif N-terminal of the FHA domain.

- An ORF from the bacterium Streptomyces, which is on the opposite strand of the protein
kinase pks1, overlapping the ORF of the kinase.

[1] Hofmann K.O., Bucher P. Trends Biochem. Sci. 20:347-349(1995).

[2] Stone J.M., Collinge M.A., Smith R.D., Horn M.A., Walker J.C. Science 266:793-795(1994).

5 [3] Navas T.A., Zhou Z., Elledge S.J. Cell 80:29-39(1995).

797. Ald_Xan_dh_C

Aldehyde oxidase and xanthine dehydrogenase, C terminus

10 [1] Romao MJ, Archer M, Moura I, Moura JJ, LeGall J, Engh R, Schneider M, Hof P, Huber R; Medline: 96072968 "Crystal structure of the xanthine oxidase-related aldehyde oxidoreductase from *D. gigas*." Science 1995;270:1170-1176.

Number of members: 54

798. Glyco_hydro_38

Glycosyl hydrolases family 38

Glycosyl hydrolases are key enzymes of carbohydrate metabolism.

20 Number of members: 20

[1] Henrissat B; Medline: 98313424; "Glycosidase families" Biochem Soc Trans 1998;26:153-156.

25 799. HECT

HECT-domain (ubiquitin-transferase).

The name HECT comes from Homologous to the E6-AP Carboxyl Terminus.

30 Number of members: 43

662

[1] Huibregtse JM, Scheffner M, Beaudenon S, Howley PM; Medline: 95223981; "A family of proteins structurally and functionally related to the E6-AP ubiquitin-protein ligase." Proc Natl Acad Sci U S A 1995;92:2563-2567.

5 800. HRDC

HRDC domain

The HRDC (Helicase and RNase D C-terminal) domain has a putative role in nucleic acid binding. Mutations in the HRDC domain cause human disease.

10 Number of members: 19

[1] Morozov V, Mushegian AR, Koonin EV, Bork P; Medline: 98060076; "A putative nucleic acid-binding domain in Bloom's and Werner's syndrome helicases" Trends Biochem Sci 1997;22:417-418.

5 801. Integrase

Integrase mediates integration of a DNA copy of the viral genome into the host chromosome. Integrase is composed of three domains. The amino-terminal domain is a zinc binding domain. The central domain is the catalytic domain [1]. The carboxyl terminal domain is a DNA binding domain [2].

20 Number of members: 581

[1] Dyda F, Hickman AB, Jenkins TM, Engelman A, Craigie R, Davies DR; Medline: 95099322. "Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases." Science 1994;266:1981-1986.

[2] Lodi PJ, Ernst JA, Kuszewski J, Hickman AB, Engelman A, Craigie R, Clore GM, Gronenborn AM; Medline: 95359147; "Solution structure of the DNA binding domain of HIV-1 integrase." Biochemistry 1995;34:9826-9833

30 802. lig_chan

Ligand-gated ion channel

663

This family includes the four transmembrane regions of the ionotropic glutamate receptors and NMDA receptors.

Number of members: 128

5

[1] Tong G, Shepherd D, Jahr CE; Medline: 95184014; "Synaptic desensitization of NMDA receptors by calcineurin." Science 1995;267:1510-1512.

803. RhoGAP

10 RhoGAP domain

GTPase activator proteins towards Rho/Rac/Cdc42-like small GTPases.

Number of members: 97

15

[1] Musacchio A, Cantley LC, Harrison SC; Medline: 97121392; "Crystal structure of the breakpoint cluster region-homology domain from phosphoinositide 3-kinase p85 alpha subunit." Proc Natl Acad Sci U S A 1996;93:14373-14378.

20

[2] Barrett T, Xiao B, Dodson EJ, Dodson G, Ludbrook SB, Nurmahomed K, Gamblin SJ, Musacchio A, Smerdon SJ, Eccleston JF; Medline: 97162209; "The structure of the GTPase-activating domain from p50rhoGAP." Nature 1997;385:458-461.

[3] Rittinger K, Walker PA, Eccleston JF, Nurmahomed K, Owen D, Laue E, Gamblin SJ, Smerdon SJ; Medline: 97404320; "Crystal structure of a small G protein in complex with the GTPase-activating protein rhoGAP." Nature 1997;388:693-697.

25

[4] Boguski MS, McCormick F; Medline: 94081948; "Proteins regulating Ras and its relatives." Nature 1993;366:643-654.

804. vwd

von Willebrand factor type D domain

30

[1] Bork P; Medline: 93327926; "The modular architecture of a new family of growth regulators related to connective tissue growth factor." FEBS lett 1993;327:125-130.

Number of members: 92

805. zf-C4_Topoiso

Topoisomerase DNA binding C4 zinc finger

- 5 [1] Tse-Dinh YC, Beran-Steed RK; Medline: 89034032; "Escherichia coli DNA topoisomerase I is a zinc metalloprotein with three repetitive zinc-binding domains." J Biol Chem 1988;263:15857-15859.
- [2] Ahumada A, Tse-Dinh YC; Medline: 99011409; "The Zn(II) binding motifs of E. coli DNA topoisomerase I is part of a high-affinity DNA binding domain." Biochem Biophys Res Commun 1998;251:509-514.
- 10

Number of members: 51

806. AIRC

AIR carboxylase

Members of this family catalyse the decarboxylation of 1-(5-phosphoribosyl)-5-amino-4-imidazole-carboxylate (AIR). This family catalyse the sixth step of de novo purine biosynthesis. Some members of this family contain two copies of this domain. Number of members: 35

15

20

807. Bromodomain signature and profile

PROSITE cross-reference(s): PS00633; BROMODOMAIN_1, PS50014;

BROMODOMAIN_2

- 25 The bromodomain [1,2,3] is a conserved region of about 70 amino acids found in the following proteins:

- Higher eukaryotes transcription initiation factor TFIID 250 Kd subunit (TBP-associated factor p250) (gene CCG1). P250 associated with the TFIID TATA-box binding protein and seems essential for progression of the G1 phase of the cell cycle.

30

- Human RING3, a protein of unknown function encoded in the MHC class II locus.

- Mammalian CREB-binding protein (CBP), which mediates cAMP-gene regulation by binding specifically to phosphorylated CREB protein.

- *Drosophila* female sterile homeotic protein (gene *fish*), required maternally for proper expression of other homeotic genes involved in pattern formation, such as *Ubx*.

- *Drosophila* *brahma* protein (gene *brm*), a protein required for the activation of multiple homeotic genes.

5 - Mammalian homologs of *brahma*. In human, three *brahma*-like proteins are known: SNF2a(hBRM), SNF2b, and BRG1.

- Human BS69, a protein that binds to adenovirus E1A and inhibits E1A transactivation

- Human peregrin (or Br140).

- Yeast BDF1 [3], a transcription factor involved in the expression of a broad class of genes
10 including snRNAs.

- Yeast GCN5, a general transcriptional activator operating in concert with certain other DNA-binding transcriptional activators, such as GCN4, HAP2/3/4 or ADA2.

- Yeast NPS1/STH1, involved in G(2) phase control in mitosis.

- Yeast SNF2/SWI2, which is part of a complex with the SNF5, SNF6, SWI3 and ADR6/SWI1 proteins. This SWI-complex is involved in transcriptional activation.

- Yeast SPT7, a transcriptional activator of Ty elements and possibly other genes.

- *Caenorhabditis elegans* protein *cbp-1*.

- Yeast hypothetical protein YGR056w.

- Yeast hypothetical protein YKR008w.

- Yeast hypothetical protein L9638.1.
15

Some proteins contain a region which, while similar to some extent to a classical bromodomain, diverges from it by either lacking part of the domain or because of an insertion. These proteins are:

25 - Mammalian protein HRX (also known as All-1 or MLL), a protein involved in translocations leading to acute leukemias and which possibly acts as a transcriptional regulatory factor. HRX contains a region similar to the C- terminal half of the bromodomain.

- *Caenorhabditis elegans* hypothetical protein ZK783.4. The bromodomain of this protein has
30 a 23 amino-acid insertion.

- Yeast protein YTA7. This protein contains a region with significant similarity to the C-terminal half of the bromodomain. As it is a member of the AAA family (see <PDOC00572>) it is also in a functionally different context.

The above proteins generally contain a single bromodomain, but some of them contain two copies, this is the case of BDF1, CCG1, fsh, RING3, YKR008w and L9638.1.

- 5 The exact function of this domain is not yet known but it is thought to be involved in protein-protein interactions and it may be important for the assembly or activity of multicomponent complexes involved in transcriptional activation.

10 The consensus pattern that has been developed spans a major part of the bromodomain; a more sensitive detection is available through the use of a profile which spans the whole domain.

Consensus pattern[STANVF]-x(2)-F-x(4)-[DNS]-x(5,7)-[DENQTF]-Y-[HFY]-x(2)-
[LIVMFY]-x(3)-[LIVM]-x(4)-[LIVM]-x(6,8)-Y-x(12,13)-[LIVM]-
5 x(2)-N-[SACF]-x(2)-[FY]

References

- [1] Haynes S.R., Doolard C., Winston F., Beck S., Trowsdale J., Dawid I.B. Nucleic Acids Res. 20:2693-2603(1992).
20 [2] Tamkun J.W., Deuring R., Scott M.P., Kissinger M., Pattatucci A.M., Kaufman T.C., Kennison J.A. Cell 68:561-572(1992).
[3] Tamkun J.W. Curr. Opin. Genet. Dev. 5:473-477(1995).

808. (CH) Actinin-type actin-binding domain signatures

25 PROSITE cross-reference(s): PS00019; ACTININ_1, PS00020; ACTININ_2

Alpha-actinin is a F-actin cross-linking protein which is thought to anchor actin to a variety of intracellular structures [1]. The actin-binding domain of alpha-actinin seems to reside in the first 250 residues of the protein. A similar actin-binding domain has been found in the N-
30 terminal region of many different actin-binding proteins [2,3]:

- In the beta chain of spectrin (or fodrin).

- In dystrophin, the protein defective in Duchenne muscular dystrophy (DMD) and which may play a role in anchoring the cytoskeleton to the plasma membrane.

- In the slime mold gelation factor (or ABP-120).

- In actin-binding protein ABP-280 (or filamin), a protein that link actin filaments to membrane glycoproteins.

- In fimbrin (or plastin), an actin-bundling protein. Fimbrin differs from the above proteins in that it contains two tandem copies of the actin-binding domain and that these copies are located in the C-terminal part of the protein.

Two conserved regions were selected as signature patterns for this type of main. The first of this region is located at the beginning of the domain, hile the second one is located in the central section and has been shown to be essential for the binding of actin.

Consensus pattern[EQ]-x(2)-[ATV]-[FY]-x(2)-W-x-N

Consensus pattern[LIVM]-x-[SGN]-[LIVM]-[DAGHE]-[SAG]-x-[DNEAG]-[LIVM]-x-[DEAG]-x(4)-[LIVM]-x-[LM]-[SAG]-[LIVM]-[LIVMT]-W-x- [LIVM](2)

[1] Schleicher M., Andre E., Harmann A., Noegel A.A. Dev. Genet. 9:521-530(1988).

[2] Matsudaira P. Trends Biochem. Sci. 16:87-92(1991).

[3] Dubreuil R.R. BioEssays 13:219-226(1991).

809. (COX1) Heme-copper oxidase subunit I, copper B binding region signature

PROSITE cross-reference(s): PS00077; COX1

Heme-copper respiratory oxidases [1] are oligomeric integral membrane protein complexes that catalyze the terminal step in the respiratory chain: they transfer electrons from cytochrome c or a quinol to oxygen. Some terminal oxidases generate a transmembrane proton gradient across the plasma membrane (prokaryotes) or the mitochondrial inner membrane (eukaryotes). The enzyme

complex consists of 3-4 subunits (prokaryotes) up to 13 polypeptides (mammals)

of which only the catalytic subunit (equivalent to mammalian subunit 1 (CO I)) is found in all heme-copper respiratory oxidases. The presence of a bimetallic

center (formed by a high-spin heme and copper B) as well as a low-spin heme, both ligated to six conserved histidine residues near the outer side of four

transmembrane spans within CO I is common to all family members [2-4].

In contrary to eukaryotes the respiratory chain of prokaryotes is branched to multiple terminal oxidases. The enzyme complexes vary in heme and copper composition, substrate type and substrate affinity. The different respiratory oxidases allow the cells to customize their respiratory systems according a variety of environmental growth conditions [1].

Recently also a component of an anaerobic respiratory chain has been found to contain the copper B binding signature of this family: nitric oxide reductase (NOR) exists in denitrifying species of Archae and Eubacteria.

Enzymes that belong to this family are:

- Mitochondrial-type cytochrome c oxidase (EC 1.9.3.1) which uses cytochrome c as electron donor. The electrons are transferred via copper A (Cu(A)) and heme a to the bimetallic center of CO I that is formed by a penta-coordinated heme a and copper B (Cu(B)). Subunit 1 contains 12 transmembrane regions. Cu(B) is said to be ligated to three of the conserved histidine residues within the transmembrane segments 6 and 7.
- Quinol oxidase from prokaryotes that transfers electrons from a quinol to the binuclear center of polypeptide I. This category of enzymes includes *Escherichia coli* cytochrome O terminal oxidase complex which is a component of the aerobic respiratory chain that predominates when cells are grown at high aeration.
- FixN, the catalytic subunit of a cytochrome c oxidase expressed in nitrogen-fixing bacteroids living in root nodules. The high affinity for oxygen allows oxidative phosphorylation under low oxygen concentrations. A similar enzyme has been found in other purple bacteria.
- Nitric oxide reductase (EC 1.7.99.7) from *Pseudomonas stutzeri*. NOR reduces nitrate to dinitrogen. It is a heterodimer of norC and the catalytic subunit norB. The latter contains the 6 invariant histidine residues and 12 transmembrane segments [5].

As a signature pattern the copper-binding region was used.

Consensus pattern[YWG]-[LIVFYWTA](2)-[VGS]-H-[LNP]-x-V-x(44,47)-H-H [The
5 three H's are copper B ligands]

Notecytochrome bd complexes do not belong to this family.

[1]

10 Garcia-Horsman J.A., Barquera B., Rumbley J., Ma J., Gennis R.B.
J. Bacteriol. 176:5587-5600(1994).

[2]

Castresana J., Luebben M., Saraste M., Higgins D.G.
EMBO J. 13:2516-2525(1994).

[3]

Capaldi R.A., Malatesta F., Darley-USmar V.M.
Biochim. Biophys. Acta 726:135-148(1983).

[4]

Holm L., Saraste M., Wikstrom M.
EMBO J. 6:2819-2823(1987).

[5]

Saraste M., Castresana J.
FEBS Lett. 341:1-4(1994).

25 810. (dehydrog_molyb) Eukaryotic molybdopterin oxidoreductases signature
PROSITE cross-reference(s): PS00559; MOLYBDOPTERIN_EUK

A number of different eukaryotic oxidoreductases that require and bind a
molybdopterin cofactor have been shown [1] to share a few regions of sequence
30 similarity. These enzymes are:

- Xanthine dehydrogenase (EC 1.1.1.204), which catalyzes the oxidation of
xanthine to uric acid with the concomitant reduction of NAD. Structurally,

670

this enzyme of about 1300 amino acids consists of at least three distinct domains: an N-terminal 2Fe-2S ferredoxin-like iron-sulfur binding domain (see <PDOC00175>), a central FAD/NAD-binding domain and a C-terminal Mo-pterin domain.

5 - Aldehyde oxidase (EC 1.2.3.1), which catalyzes the oxidation aldehydes into acids. Aldehyde oxidase is highly similar to xanthine dehydrogenase in its sequence and domain structure.

- Nitrate reductase (EC 1.6.6.1), which catalyzes the reduction of nitrate to nitrite. Structurally, this enzyme of about 900 amino acids consists of
10 an N-terminal Mo-pterin domain, a central cytochrome b5-type heme-binding domain (see <PDOC00170>) and a C-terminal FAD/NAD-binding cytochrome reductase domain.

- Sulfite oxidase (EC 1.8.3.1), which catalyzes the oxidation of sulfite to sulfate. Structurally, this enzyme of about 460 amino acids consists of an
15 N-terminal cytochrome b5-binding domain followed by a Mo-pterin domain.

There are a few conserved regions in the sequence of the molybdopterin-binding domain of these enzymes. The pattern used to detect these proteins is based on one of them. It contains a cysteine residue which could be involved in
20 binding the molybdopterin cofactor.

Consensus pattern[GA]-x(3)-[KRNQHT]-x(11,14)-[LIVMFYWS]-x(8)-[LIVMF]-x-C-x(2)-[DEN]-R-x(2)-[DE]

25 [1]

Wootton J.C., Nicolson R.E., Cock J.M., Walters D.E., Burke J.F., Doyle W.A., Bray R.C.
Biochim. Biophys. Acta 1057:157-185(1991).

30 811. (DNA_ligase) ATP-dependent DNA ligase signatures

PROSITE cross-reference(s): PS00697; DNA_LIGASE_A1, PS00333; DNA_LIGASE_A2

DNA ligase (polydeoxyribonucleotide synthase) is the enzyme that joins two DNA

671

fragments by catalyzing the formation of an internucleotide ester bond between phosphate and deoxyribose. It is active during DNA replication, DNA repair and DNA recombination. There are two forms of DNA ligase: one requires ATP (EC 6.5.1.1), the other NAD (EC 6.5.1.2).

5 Eukaryotic, archaebacterial, virus and phage DNA ligases are ATP-dependent. During the first step of the joining reaction, the ligase interacts with ATP to form a covalent enzyme-adenylate intermediate. A conserved lysine residue is the site of adenylation [1,2].

10 Apart from the active site region, the only conserved region common to all ATP-dependent DNA ligases is found [3] in the C-terminal section and contains a conserved glutamate as well as four positions with conserved basic residues.

15 Signature patterns were developed for both conserved regions.

Consensus pattern[EDQH]-x-K-x-[DN]-G-x-R-[GACIVM] [K is the active site residue]

20 Consensus patternE-G-[LIVMA]-[LIVM](2)-[KR]-x(5,8)-[YW]-[QNEK]-x(2,6)-[KRH]-x(3,5)-K-[LIVMFY]-K

Sequences known to belong to this class detected by the patternALL, except for archebacterial DNA ligases.

25 [1]
Tomkinson A.E., Totty N.F., Ginsburg M., Lindahl T.
Proc. Natl. Acad. Sci. U.S.A. 88:400-404(1991).

[2]
Lindahl T., Barnes D.E.

30 Annu. Rev. Biochem. 61:251-281(1992).

[3]
Kletzin A.
Nucleic Acids Res. 20:5389-5396(1992).

812. (FAD_Gly3P_dh) FAD-dependent glycerol-3-phosphate dehydrogenase signatures
 PROSITE cross-reference(s): PS00977; FAD_G3PDH_1, PS00978; FAD_G3PDH_2

5 FAD-dependent glycerol-3-phosphate dehydrogenase (EC 1.1.99.5) (GPD) catalyzes
 the conversion of glycerol-3-phosphate into dihydroxyacetone phosphate. In
 bacteria [1] it is associated with the utilization of glycerol coupled to
 respiration. In *Escherichia coli*, two isozymes are known: one expressed under
 anaerobic conditions (gene *glpA*) and one in aerobic conditions (gene *glpD*). In
 10 eukaryotes, a mitochondrial form of GPD participates in the glycerol phosphate
 shuttle in conjunction with an NAD-dependent cytoplasmic GPD (EC 1.1.1.8) [2,
 3].

These enzymes are proteins of about 60 to 70 Kd which contain a probable
 15 FAD-binding domain in their N-terminal extremity. The mammalian enzyme differs
 from the bacterial or yeast proteins by having an EF-hand calcium-binding
 region (See <PDOC00018>) in its C-terminal extremity.

Two signature patterns were developed. One based on the first half of the FAD-
 20 binding domain and one which corresponds to a conserved region in the central
 part of these enzymes.

Consensus pattern[IV]-G-G-G-x(2)-G-[STACV]-G-x-A-x-D-x(3)-R-G

25 Consensus patternG-G-K-x(2)-[GSTE]-Y-R-x(2)-A

[1]

Austin D., Larson T.J.

J. Bacteriol. 173:101-107(1991).

[2]

30 Roennow B., Kielland-Brandt M.C.

Yeast 9:1121-1130(1993).

[3]

Brown L.J., McDonald M.J., Lehn D.A., Moran S.M.

J. Biol. Chem. 269:14363-14366(1994).

813. (Fapy_DNA_glyco) Formamidopyrimidine-DNA glycosylase signature
PROSITE cross-reference(s): PS01242; FPG

5

Formamidopyrimidine-DNA glycosylase (EC 3.2.2.23) [1] (Fapy-DNA glycosylase) (gene fpg) is a bacterial enzyme involved in DNA repair and which excise oxidized purine bases to release 2,6-diamino-4-hydroxy-5N-methylformamido-pyrimidine (Fapy) and 7,8-dihydro-8-oxoguanine (8-OxoG) residues. In addition to its glycosylase activity, FPG can also nick DNA at apurinic/apyrimidinic sites (AP sites). FPG is a monomeric protein of about 32 Kd which binds and require zinc for its activity.

10

The binding site for zinc seems to be located in the C-terminal part of the enzyme where four conserved and essential [2] cysteines are located. A signature pattern was developed based on this region.

15

Consensus pattern C-x(2,4)-C-x-[GTAQ]-x-[IV]-x(7)-R-[GSTAN]-[STA]-x-[FYI]-C-x(2)-C-Q

20

[The four C's are putative zinc ligands]

[1]

Duwat P., de Oliveira R., Ehrlich S.D., Boiteux S.

Microbiology 141:411-417(1995).

25

[2]

O'Connor T.E., Graves R.J., Demurcia G., Castaing B., Laval J.

J. Biol. Chem. 268:9063-9070(1993).

814. (G_glu_transpept) Gamma-glutamyltranspeptidase signature

30

PROSITE cross-reference(s): PS00462; G_GLU_TRANSPEPTIDASE

Gamma-glutamyltranspeptidase (EC 2.3.2.2) (GGT) [1] catalyzes the transfer of the gamma-glutamyl moiety of glutathione to an acceptor that may be an amino

674

acid, a peptide or water (forming glutamate). GGT plays a key role in the gamma-glutamyl cycle, a pathway for the synthesis and degradation of glutathione. In prokaryotes and eukaryotes, it is an enzyme that consists of two polypeptide chains, a heavy and a light subunit, processed from a single chain precursor. The active site of GGT is known to be located in the light subunit.

The sequences of mammalian and bacterial GGT show a number of regions of high similarity [2]. *Pseudomonas cephalosporin acylases* (EC 3.5.1.-) that convert 7-beta-(4-carboxybutanamido)-cephalosporanic acid (GL-7ACA) into 7-aminocephalosporanic acid (7ACA) and glutaric acid are evolutionary related to GGT and also show some GGT activity [3]. Like GGT, these GL-7ACA acylases, are also composed of two subunits.

One of the conserved regions correspond to the N-terminal extremity of the mature light chains of these enzymes. This region was used as a signature pattern.

Consensus pattern T-[STA]-H-x-[ST]-[LIVMA]-x(4)-G-[SN]-x-V-[STA]-x-T-x-T-[LIVM]-[NE]-x(1,2)-[FY]-G

[1]

Tate S.S., Meister A.

Meth. Enzymol. 113:400-419(1985).

[2]

Suzuki H., Kumagai H., Echigo T., Tochikura T.

J. Bacteriol. 171:5169-5172(1989).

[3]

Ishiye M., Niwa M.

Biochim. Biophys. Acta 1132:233-239(1992).

815. G-protein gamma subunit profile

PROSITE cross-reference(s): PS50058; G_PROTEIN_GAMMA

Guanine nucleotide-binding proteins (G proteins) [1] act as intermediaries in the transduction of signals generated by transmembrane receptors. G proteins consist of three subunits (alpha, beta, and gamma). The alpha subunit binds to and hydrolyzes GTP; the functions of the beta and gamma subunits are less clear but they seem to be required for the replacement of GDP by GTP as well as for membrane anchoring and receptor recognition.

The gamma subunits are small proteins (from 70 to 110 residues) that are bound to the membrane via a isoprenyl group (either a farnesyl or a geranyl-geranyl) covalently linked to their C-terminus. In mammals there are at least 12 different isoforms of gamma subunits.

The *Caenorhabditis elegans* protein egl-10, which is a regulator of G-protein signalling, contains a G-protein gamma-like domain.

A profile was developed that spans the complete length of the gamma subunit.

[1]

Pennington S.R.

Protein Prof. 2:16-315(1995).

816. GNS1/SUR4 family signature

PROSITE cross-reference(s): PS01188; GNS1_SUR4

The following group of eukaryotic integral membrane proteins, whose exact function has not yet clearly been established, are evolutionary related [1]:

- Yeast GNS1 [2], a protein involved in synthesis of 1,3-beta-glucan.
- Yeast SUR4 (or APA1, SRE1) [3], a protein that could act in a glucose-signaling pathway that controls the expression of several genes that are transcriptionally regulated by glucose.

- Yeast hypothetical protein YJL196c.
- Caenorhabditis elegans hypothetical protein C40H1.4.
- Caenorhabditis elegans hypothetical protein D2024.3.

5 The proteins have from 290 to 435 amino acid residues. Structurally, they seem to be formed of three sections: a N-terminal region with two transmembrane domains, a central hydrophilic loop and a C-terminal region that contains from one to three transmembrane domains. A conserved region that contains three histidines was selected as a signature pattern. This region is located in the
10 hydrophilic loop.

Consensus pattern L-x-F-L-H-x-Y-H-H

[1]

15 Bairoch A.

Unpublished observations (1996).

[2]

El-Sherbeini M., Clemas J.A.

J. Bacteriol. 177:3227-3234(1995).

20 [3]

Garcia-Arranz M., Maldonado A.M., Mazon M.J., Portillo F.

J. Biol. Chem. 269:18076-18082(1994).

817. Immunoglobulins and major histocompatibility complex proteins signature

25 PROSITE cross-reference(s): PS00290; IG_MHC

The basic structure of immunoglobulin (Ig) [1] molecules is a tetramer of two light chains and two heavy chains linked by disulfide bonds. There are two types of light chains: kappa and lambda, each composed of a constant domain (CL) and a variable domain (VL). There are five types of heavy chains: alpha, delta, epsilon, gamma and mu, all consisting of a variable domain (VH) and three (in alpha, delta and gamma) or four (in epsilon and mu) constant domains (CH1 to CH4).
30

The major histocompatibility complex (MHC) molecules are made of two chains.

In class I [2] the alpha chain is composed of three extracellular domains, a transmembrane region and a cytoplasmic tail. The beta chain (beta-2-

5 microglobulin) is composed of a single extracellular domain. In class II [3], both the alpha and the beta chains are composed of two extracellular domains, a transmembrane region and a cytoplasmic tail.

It is known [4,5] that the Ig constant chain domains and a single

10 extracellular domain in each type of MHC chains are related. These homologous domains are approximately one hundred amino acids long and include a conserved intradomain disulfide bond. A small pattern around the C-terminal cysteine is involved in this disulfide bond which can be used to detect these category of Ig related proteins.

Consensus pattern[FY]-x-C-x-[VA]-x-H-Sequences known to belong to this

class detected by the pattern: Ig heavy chains type Alpha C region : All,

in CH2 and CH3. Ig heavy chains type Delta C region : All, in CH3. Ig

heavy chains type Epsilon C region: All, in CH1, CH3 and CH4. Ig heavy

20 chains type Gamma C region : All, in CH3 and also CH1 in some cases Ig

heavy chains type Mu C region : All, in CH2, CH3 and CH4. Ig light chains

type Kappa C region : In all CL except rabbit and Xenopus. Ig light chains

type Lambda C region : In all CL except rabbit. MHC class I alpha chains :

All, in alpha-3 domains, including in the cytomegalovirus MHC-1 homologous

25 protein [6]. Beta-2-microglobulin : All. MHC class II alpha chains: All,

in alpha-2 domains. MHC class II beta chains: All, in beta-2 domains.

[1]

Gough N.

30 Trends Biochem. Sci. 6:203-205(1981).

[2]

Klein J., Figueroa F.

Immunol. Today 7:41-44(1986).

[3]

Figuerola F., Klein J.

Immunol. Today 7:78-81(1986).

[4]

5 Orr H.T., Lancet D., Robb R.J., Lopez de Castro J.A., Strominger J.L.

Nature 282:266-270(1979).

[5]

Cushley W., Owen M.J.

Immunol. Today 4:88-92(1983).

10 [6]

Beck S., Barrel B.G.

Nature 331:269-272(1988).

818. (IGFBP) Insulin-like growth factor binding proteins signature

15 PROSITE cross-reference(s): PS00222; IGF_BINDING

The insulin-like growth factors (IGF-I and IGF-II) bind to specific binding proteins in extracellular fluids with high affinity [1,2,3]. These IGF-binding proteins (IGFBP) prolong the half-life of the IGFs and have been shown to either inhibit or stimulate the growth promoting effects of the IGFs on cells culture. They seem to alter the interaction of IGFs with their cell surface receptors. There are at least six different IGFBPs and they are structurally related.

25 The following growth-factor inducible proteins are structurally related to IGFBPs and could function as growth-factor binding proteins [4,5]:

- Mouse protein cyr61 and its probable chicken homolog, protein CEF-10.
- Human connective tissue growth factor (CTGF) and its mouse homolog, protein
- 30 FISP-12.
- Vertebrate protein NOV.

As a signature pattern a conserved cysteine-rich region located in the N-terminal

section of these proteins is used.

Consensus pattern G-C-[GS]-C-C-x(2)-C-A-x(6)-C

Sequences known to belong to this class detected by the pattern ALL, except
5 for IGFBP-6's.

[1]

Rechler M.M.

Vitam. Horm. 47:1-114(1993).

10 [2]

Shimasaki S., Ling N.

Prog. Growth Factor Res. 3:243-266(1991).

[3]

Clemmons D.R.

15 Trends Endocrinol. Metab. 1:412-417(1990).

[4]

Bradham D.M., Igarashi A., Potter R.L., Grotendorst G.R.

J. Cell Biol. 114:1285-1294(1991).

[5]

20 Maloisel V., Martinerie C., Dambrine G., Plassiart G., Brisac M., Crochet

J., Perbal B.

Mol. Cell. Biol. 12:10-21(1992).

819. LMWPc : Low molecular weight phosphotyrosine protein phosphatase

25 Number of members: 34

[1]Medline: 94329182, The crystal structure of a low-molecular-weight phosphotyrosine
protein phosphatase. Su XD, Taddei N, Stefani M, Ramponi G, Nordlund P; Nature
1994;370:575-578.

30

820. (myosin_head) ATP/GTP-binding site motif A (P-loop)

PROSITE cross-reference(s): PS00017; ATP_GTP_A

From sequence comparisons and crystallographic data analysis it has been shown [1,2,3,4,5,6] that an appreciable proportion of proteins that bind ATP or GTP share a number of more or less conserved sequence motifs. The best conserved of these motifs is a glycine-rich region, which typically forms a flexible

5 loop between a beta-strand and an alpha-helix. This loop interacts with one of the phosphate groups of the nucleotide. This sequence motif is generally referred to as the 'A' consensus sequence [1] or the 'P-loop' [5].

There are numerous ATP- or GTP-binding proteins in which the P-loop is found.

10 A number of protein families for which the relevance of the presence of such motif has been noted is listed below:

- ATP synthase alpha and beta subunits (see <PDOC00137>).
- Myosin heavy chains.
- 15 - Kinesin heavy chains and kinesin-like proteins (see <PDOC00343>).
- Dynamins and dynamin-like proteins (see <PDOC00362>).
- Guanylate kinase (see <PDOC00670>).
- Thymidine kinase (see <PDOC00524>).
- Thymidylate kinase (see <PDOC01034>).
- 20 - Shikimate kinase (see <PDOC00868>).
- Nitrogenase iron protein family (nifH/frxC) (see <PDOC00580>).
- ATP-binding proteins involved in 'active transport' (ABC transporters) [7] (see <PDOC00185>).
- DNA and RNA helicases [8,9,10].
- 25 - GTP-binding elongation factors (EF-Tu, EF-1alpha, EF-G, EF-2, etc.).
- Ras family of GTP-binding proteins (Ras, Rho, Rab, Ral, Ypt1, SEC4, etc.).
- Nuclear protein ran (see <PDOC00859>).
- ADP-ribosylation factors family (see <PDOC00781>).
- Bacterial dnaA protein (see <PDOC00771>).
- 30 - Bacterial recA protein (see <PDOC00131>).
- Bacterial recF protein (see <PDOC00539>).
- Guanine nucleotide-binding proteins alpha subunits (Gi, Gs, Gt, G0, etc.).
- DNA mismatch repair proteins mutS family (See <PDOC00388>).

- Bacterial type II secretion system protein E (see <PDOC00567>).

Not all ATP- or GTP-binding proteins are picked-up by this motif. A number of proteins escape detection because the structure of their ATP-binding site is completely different from that of the P-loop. Examples of such proteins are the E1-E2 ATPases or the glycolytic kinases. In other ATP- or GTP-binding proteins the flexible loop exists in a slightly different form; this is the case for tubulins or protein kinases. A special mention must be reserved for adenylate kinase, in which there is a single deviation from the P-loop pattern: in the last position Gly is found instead of Ser or Thr.

Consensus pattern[AG]-x(4)-G-K-[ST]

[1]

Walker J.E., Saraste M., Runswick M.J., Gay N.J.
EMBO J. 1:945-951(1982).

[2]

Moller W., Amons R.
FEBS Lett. 186:1-7(1985).

[3]

Fry D.C., Kuby S.A., Mildvan A.S.
Proc. Natl. Acad. Sci. U.S.A. 83:907-911(1986).

[4]

Dever T.E., Glynias M.J., Merrick W.C.
Proc. Natl. Acad. Sci. U.S.A. 84:1814-1818(1987).

[5]

Saraste M., Sibbald P.R., Wittinghofer A.
Trends Biochem. Sci. 15:430-434(1990).

[6]

Koonin E.V.
J. Mol. Biol. 229:1165-1174(1993).

[7]

Higgins C.F., Hyde S.C., Mimmack M.M., Gileadi U., Gill D.R., Gallagher

M.P.

J. Bioenerg. Biomembr. 22:571-592(1990).

[8]

Hodgman T.C.

5 Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).

[9]

Linder P., Lasko P., Ashburner M., Leroy P., Nielsen P.J., Nishi K.,

Schnier J., Slonimski P.P.

Nature 337:121-122(1989).

10 [10]

Gorbalenya A.E., Koonin E.V., Donchenko A.P., Blinov V.M.

Nucleic Acids Res. 17:4713-4730(1989).

821. PE: PE family

15 This family named after a PE motif near to the amino terminus of the domain. The PE family of proteins all contain an amino-terminal region of about 110 amino acids. The carboxyl terminus of this family are variable and fall into several classes. The largest class of PE proteins is the highly repetitive PGRS class which have a high glycine content. The function of these proteins is uncertain but it has been suggested that they may be related to antigenic variation of Mycobacterium tuberculosis [1]. Number of members: 88

20 [1] Medline: 98295987. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Barrell BG, et al; Nature 1998;393:537-544.

822. (RNB) Ribonuclease II family signature

PROSITE cross-reference(s): PS01175; RIBONUCLEASE_II

30

On the basis of sequence similarities, the following bacterial and eukaryotic proteins seem to form a family:

- *Escherichia coli* and related bacteria ribonuclease II (EC 3.1.13.1) (RNase II) (gene *rnb*) [1]. RNase II is an exonuclease involved in mRNA decay. It degrades mRNA by hydrolyzing single-stranded polyribonucleotides processively in the 3' to 5' direction.

- 5 - Bacterial protein *vacB*. In *Shigella flexneri*, *vacB* has been shown to be required for the expression of virulence genes at the posttranscriptional level.
- Yeast protein SSD1 (or SRK1) which is implicated in the control of the cell cycle G1 phase.
- 10 - Yeast protein DIS3 [2], which binds to *ran* (GSP1) and enhances the nucleotide-releasing activity of RCC1 on *ran*.
- Fission yeast protein *dis3*, which is implicated in mitotic control.
- *Neurospora crassa* *cyt-4*, a mitochondrial protein required for RNA 5' and 3' end processing and splicing.
- 15 - Yeast protein MSU1, which is involved in mitochondrial biogenesis.
- *Synechocystis* strain PCC 6803 protein *zam* [3], which control resistance to the carbonic anhydrase inhibitor acetazolamide.
- *Caenorhabditis elegans* hypothetical protein F48E8.6.

20 The size of these proteins range from 644 residues (*rnb*) to 1250 (SSD1). While their sequence is highly divergent they share a conserved domain in their C-terminal section [4]. It is possible that this domain plays a role in a putative exonuclease function that would be common to all these proteins. A signature pattern was developed based on the core of this conserved domain.

25 Consensus pattern[HI]-[FYE]-[GSTAM]-[LIVM]-x(4,5)-Y-[STAL]-x-[FWVAC]-[TV]-[SA]-P-[LIVMA]-[RQ]-[KR]-[FY]-x-D-x(3)-[HQ]

[1]

30 Zilhao R., Camelo L., Arraiano C.M.
Mol. Microbiol. 8:43-51(1993).

[2]

Noguchi E., Hayashi N., Azuma Y., Seki T., Nakamura M., Nakashima N.,

Yanagida M., He X., Mueller U., Sazer S., Nishimoto T.

EMBO J. 15:5595-5605(1996).

[3]

Beuf L., Bedu S., Cami B., Joset F.

5 Plant Mol. Biol. 27:779-788(1995).

[4]

Mian I.S.

Nucleic Acids Res. 25:3187-3195(1997).

10 823. Src homology 2 (SH2) domain profile

PROSITE cross-reference(s): PS50001; SH2

The Src homology 2 (SH2) domain is a protein domain of about 100 amino-acid residues first identified as a conserved sequence region between the
15 oncoproteins Src and Fps [1]. Similar sequences were later found in many other intracellular signal-transducing proteins [2]. SH2 domains function as regulatory modules of intracellular signalling cascades by interacting with high affinity to phosphotyrosine-containing target peptides in a sequence-specific and strictly phosphorylation-dependent manner [3,4,5,6].

20 The SH2 domain has a conserved 3D structure consisting of two alpha helices and six to seven beta-strands. The core of the domain is formed by a continuous beta-meander composed of two connected beta-sheets [7].

25 So far, SH2 domains have been identified in the following proteins:

- Many vertebrate, invertebrate and retroviral cytoplasmic (non-receptor) protein tyrosine kinases. In particular in the Src, Abl, Bkt, Csk and ZAP70 families of kinases.

30 - Mammalian phosphatidylinositol-specific phospholipase C gamma-1 and -2. Two copies of the SH2 domain are found in those proteins in between the catalytic 'X-' and 'Y-boxes' (see <PDOC50007>).

- Mammalian phosphatidyl inositol 3-kinase regulatory p85 subunit.

- Some vertebrate and invertebrate protein-tyrosine phosphatases.
- Mammalian Ras GTPase-activating protein (GAP).
- Adaptor proteins mediating binding of guanine nucleotide exchange factors to growth factor receptors: vertebrate GRB2, *Caenorhabditis elegans* sem-5 and *Drosophila* DRK.
- Mammalian Vav oncoprotein, a guanine-nucleotide exchange factor of the CDC24 family.
- Miscellaneous proteins interacting with vertebrate receptor protein tyrosine kinases: oncoprotein Crk, mammalian cytoplasmic proteins Nck, Shc.
- STAT proteins (signal transducers and activators of transcription).
- Chicken tensin.
- Yeast transcriptional control protein SPT6.

The profile developed to detect SH2 domains is based on a structural alignment consisting of 8 gap-free blocks and 7 linker regions totaling 92 match positions.

[1]

Sadowski I., Stone J.C., Pawson T.
Mol. Cell. Biol. 6:4396-4408(1986).

[2]

Russel R.B., Breed J., Barton G.J.
FEBS Lett. 304:15-20(1992).

[3]

Marangere L.E.M., Pawson T.
J. Cell Sci. Suppl. 18:97-104(1994).

[4]

Pawson T., Schlessinger J.
Curr. Biol. 3:434-442(1993).

[5]

Mayer B.J., Baltimore D.
Trends Cell. Biol. 3:8-13(1993).

[6]

Pawson T.

Nature 373:573-580(1995).

[7]

Kuriyan J., Cowburn D.

5 Curr. Opin. Struct. Biol. 3:828-837(1993).

824. Sulfate transporters signature

PROSITE cross-reference(s): PS01130; SULFATE_TRANSP

10 A number of proteins involved in the transport of sulfate across a membrane as well as some yet uncharacterized proteins have been shown [1,2] to be evolutionary related. These proteins are:

- Neurospora crassa sulfate permease II (gene cys-14).
- 15 - Yeast sulfate permeases (genes SUL1 and SUL2).
- Rat sulfate anion transporter 1 (SAT-1).
- Mammalian DTDST, a probable sulfate transporter which, in Human, is involved in the genetic disease, diastrophic dysplasia (DTD).
- Sulfate transporters 1, 2 and 3 from the legume Stylosanthes hamata.
- 20 - Human pendrin (gene PDS), which is involved in a number of hearing loss genetic diseases.
- Human protein DRA (Down-Regulated in Adenoma).
- Soybean early nodulin 70.
- 25 - Escherichia coli hypothetical protein ychM.
- Caenorhabditis elegans hypothetical protein F41D9.5.

As expected by their transport function, these proteins are highly hydrophobic and seem to contain about 12 transmembrane domains. The best conserved region
30 seems to be located in the second transmembrane region and is used as a signature pattern.

Consensus pattern[PAV]-x-Y-[GS]-L-Y-[STAG](2)-x(4)-[LIVFYA]-[LIVST]-[YI]-

x(3)-[GA]-[GST]-S-[KR]

[1]

Sandal N.N., Marcker K.A.

5 Trends Biochem. Sci. 19:19-19(1994).

[2]

Smith F.W., Hawkesford M.J., Prosser I.M., Clarkson D.T.

Mol. Gen. Genet. 247:709-715(1995).

10 825. TYA: TYA transposon protein

Ty are yeast transposons. A 5.7kb transcript codes for p3 a fusion protein of TYA and TYB. The TYA protein is analogous to the gag protein of retroviruses. TYA a is cleaved to form 46kd protein which can form mature virion like particles [1]. Number of members: 59

15 [1] Medline: 97404699. Cryo-electron microscopy structure of yeast Ty retrotransposon virus-like particles. Palmer KJ, Tichelaar W, Myers N, Burns NR, Butcher SJ, Kingsman AJ, Fuller SD, Saibil HR; J Virol 1997;71:6863-6868.

826. Aldolase_II

20 Class II Aldolase and Adducin N-terminal domain.

-!- This family includes class II aldolases and adducins which have not been ascribed any enzymatic function. Number of members: 37

References:

25 [1] Medline: 93294819. The spatial structure of the class II L-fucose-1-phosphate aldolase from Escherichia coli. Dreyer MK, Schulz GE; J Mol Biol 1993;231:549-553.

[2] Medline: 96256522. Catalytic mechanism of the metal-dependent fucose aldolase from Escherichia coli as derived from the structure. Dreyer MK, Schulz GE; J Mol Biol 1996;259:458-466.

30

827. CBD_2

-!- Two tryptophan residues are involved in cellulose binding.

-!- Cellulose binding domain found in bacteria. Number of members: 51

References:

[1] Medline: 95284032. Solution structure of a cellulose-binding domain from *Cellulomonas fimi* by nuclear magnetic resonance spectroscopy. Xu GY, Ong E, Gilkes NR, Kilburn DG, Muhandiram DR, Harris-Brandts M, Carver JP, Kay LE, Harvey TS; *Biochemistry* 1995;34:6993-7009.

828. P

A unique feature of the eukaryotic subtilisin-like proprotein convertases is the presence of an additional highly conserved sequence of approximately 150 residues (P domain) located immediately downstream of the catalytic domain.

Number of members: 91

References:

[1] Medline: 94252314. A C-terminal domain conserved in precursor processing proteases is required for intramolecular N-terminal maturation of pro-Kex2 protease. Gluschkof P, Fuller RS; *EMBO J* 1994;13:2280-2288.

[2] Medline: 98225190. Regulatory roles of the P domain of the subtilisin-like prohormone convertases. Zhou A, Martin S, Lipkind G, LaMendola J, Steiner DF; *J Biol Chem* 1998;273:11107-11114.

829. Uncharacterized protein family UPF0020 signature

PROSITE cross-reference(s): PS01261; UPF0020

The following uncharacterized proteins have been shown [1] to share regions of similarities:

- *Escherichia coli* hypothetical protein ycbY and HI0116/15, the corresponding *Haemophilus influenzae* protein.
- *Bacillus subtilis* hypothetical protein ypsC.
- *Synechocystis* strain PCC 6803 hypothetical protein slr0064.
- *Methanococcus jannaschii* hypothetical proteins MJ0438 and MJ0710.

These are hydrophilic proteins of from 40 Kd to about 80 Kd. They can be

picked up in the database by the following pattern.

Consensus patternD-P-[LIVMF]-C-G-[ST]-G-x(3)-[LI]-E

5 References:

[1] Bairoch A. Unpublished observations (1997).

830. Uncharacterized protein family UPF0031 signatures

PROSITE cross-reference(s): PS01049; UPF0031_1; PS01050; UPF0031_2

10 The following uncharacterized proteins have been shown [1] to share regions of similarities:

- Yeast chromosome XI hypothetical protein YKL151c.
- Caenorhabditis elegans hypothetical protein R107.2.
- 15 - Escherichia coli hypothetical protein yjeF.
- Bacillus subtilis hypothetical protein yxkO.
- Helicobacter pylori hypothetical protein HP1363.
- Mycobacterium tuberculosis hypothetical protein MtCY77.05c.
- Mycobacterium leprae hypothetical protein B229_C2_201.
- 20 - Synechocystis strain PCC 6803 hypothetical protein sll1433.
- Methanococcus jannaschii hypothetical protein MJ1586.

These are proteins of about 30 to 40 Kd whose central region is well conserved. They can be picked up in the database by the following patterns.

25

Consensus pattern[SAV]-[IVW]-[LVA]-[LIV]-G-[PNS]-G-L-[GP]-x-[DENQT]

Consensus pattern[GA]-G-x-G-D-[TV]-[LT]-[STA]-G-x-[LIVM]

831. (ACOX)

30 Acyl-CoA oxidase

This is a family of Acyl-CoA oxidases EC:1.3.3.6. Acyl-coA oxidase converts acyl-CoA into trans-2-enoyl-CoA [1].

Number of members: 39

[1] Hayashi H, De Bellis L, Yamaguchi K, Kato A, Hayashi M, Nishimura M; Medline: 98192624. "Molecular characterization of a glyoxysomal long chain acyl-CoA oxidase that is synthesized as a precursor of higher molecular mass in pumpkin." J Biol Chem 1998;273:8301-8307.

832. (AICARFT_IMPCHase)
AICARFT/IMPCHase bienzyme

This is a family of bifunctional enzymes catalysing the last steps in de novo purine biosynthesis. The bifunctional enzyme is found in both prokaryotes and eukaryotes. The second last step is catalysed by 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase EC:2.1.2.3 (AICARFT), this enzyme catalyses the formylation of AICAR with 10-formyl-tetrahydrofolate to yield FAICAR and tetrahydrofolate [1]. The last step is catalysed by IMP (Inosine monophosphate) cyclohydrolase EC:3.5.4.10 (IMPCHase), cyclizing FAICAR (5-formylaminoimidazole-4-carboxamide ribonucleotide) to IMP [1].

Number of members: 22

[1] Akira T, Komatsu M, Nango R, Tomooka A, Konaka K, Yamauchi M, Kitamura Y, Nomura S, Tsukamoto I; Medline: 97473523 "Molecular cloning and expression of a rat cDNA encoding 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase" [published erratum appears in Gene 1998 Feb 27;208(2):337] Gene 1997;197:289-293.

[2] Rayl EA, Moroson BA, Beardsley GP; Medline: 96147205 "The human purH gene product, 5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase. Cloning, sequencing, expression, purification, kinetic analysis, and domain mapping." J Biol Chem 1996;271:2225-2233.

833. (AOX)

Alternative oxidase

The alternative oxidase is used as a second terminal oxidase in the mitochondria, electrons are transferred directly from reduced ubiquinol to oxygen forming water [2]. This is not coupled to ATP synthesis and is not inhibited by cyanide, this pathway is a single step process [1]. In rice the transcript levels of the alternative oxidase are increased by low temperature [1].

Number of members: 27

[1] Ito Y, Saisho D, Nakazono M, Tsutsumi N, Hirai A; Medline: 98086211 "Transcript levels of tandem-arranged alternative oxidase genes in rice are increased by low temperature." Gene 1997;203:121-129.

[2] Li Q, Ritzel RG, McLean LL, McIntosh L, Ko T, Bertrand H, Nargang FE; Medline: 96366413 "Cloning and analysis of the alternative oxidase gene of *Neurospora crassa*." Genetics 1996;142:129-140.

834. (APH)

Protein kinases signatures and profile

Cross-reference(s): PS00107; PROTEIN_KINASE_ATP, PS00108; PROTEIN_KINASE_ST, PS00109; PROTEIN_KINASE_TYR, PS50011; PROTEIN_KINASE_DOM

Eukaryotic protein kinases [1 to 5] are enzymes that belong to a very extensive family of proteins which share a conserved catalytic core common to both serine/threonine and tyrosine protein kinases. There are a number of conserved regions in the catalytic domain of protein kinases. Two of these regions have been selected to build signature patterns. The first region, which is located in the N-terminal extremity of the catalytic domain, is a glycine-rich stretch of residues in the vicinity of a lysine residue, which has been shown to be involved in ATP

binding. The second region, which is located in the central part of the catalytic domain, contains a conserved aspartic acid residue which is important for the catalytic activity of the enzyme [6]; two signature patterns were derived for that region: one specific for serine/threonine kinases and the other for tyrosine kinases. A profile was developed which is based on the alignment in [1] and covers the entire catalytic domain.

Consensus pattern: [LIV]-G-{P}-G-{P}-[FYWMGSTNH]-[SGA]-{PW}-[LIVCAT]-{PD}-x-[GSTACLIVMFY]-x(5,18)-[LIVMFYWCSTAR]-[AIVP]-[LIVMFAGCKR]-K [K binds ATP]

Sequences known to belong to this class detected by the pattern the majority of known protein kinases but it fails to find a number of them, especially viral kinases which are quite divergent in this region and are completely missed by this pattern.

Consensus pattern: [LIVMFYC]-x-[HY]-x-D-[LIVMFY]-K-x(2)-N-[LIVMFYCT](3) [D is an active site residue]

Sequences known to belong to this class detected by the pattern. Most serine/ threonine specific protein kinases with 10 exceptions (half of them viral kinases) and also Epstein-Barr virus BGLF4 and Drosophila ninaC which have respectively Ser and Arg instead of the conserved Lys and which are therefore detected by the tyrosine kinase specific pattern described below.

Consensus pattern: [LIVMFYC]-x-[HY]-x-D-[LIVMFY]-[RSTAC]-x(2)-N-[LIVMFYC](3) [D is an active site residue] tyrosine specific protein kinases with the exception of human ERBB3 and mouse blk. This pattern will also detect most bacterial aminoglycoside phosphotransferases [8,9] and herpesviruses ganciclovir kinases [10]; which are proteins structurally and evolutionary related to protein kinases. Sequences known to belong to this class detected by the profile ALL, except for three viral kinases. This profile also detects receptor guanylate cyclases (see <PDOC00430>) and 2-5A-dependent ribonucleases. Sequence similarities between these two families and the eukaryotic protein kinase family have been noticed before. It also detects Arabidopsis thaliana kinase- like protein TMKL1 which seems to have lost its catalytic activity.

Note if a protein analyzed includes the two protein kinase signatures, the probability of it being a protein kinase is close to 100%. Note eukaryotic-type protein kinases have also been found in prokaryotes such as *Myxococcus xanthus* [11] and *Yersinia pseudotuberculosis*.

- 5 Note the patterns shown above has been updated since their publication in [7]. Note this documentation entry is linked to both signature patterns and a profile. As the profile is much more sensitive than the patterns, you should use it if you have access to the necessary software tools to do so.

10 References

- [1] Hanks S.K., Hunter T., FASEB J. 9:576-596(1995).
 [2] Hunter T., Meth. Enzymol. 200:3-37(1991).
 [3] Hanks S.K., Quinn A.M., Meth. Enzymol. 200:38-62(1991).
 [4] Hanks S.K., Curr. Opin. Struct. Biol. 1:369-383(1991).
 15 [5] Hanks S.K., Quinn A.M., Hunter T., Science 241:42-52(1988).
 [6] Knighton D.R., Zheng J., Ten Eyck L.F., Ashford V.A., Xuong N.-H., Taylor, S.S., Sowadski J.M., Science 253:407-414(1991).
 [7] Bairoch A., Claverie J.-M., Nature 331:22(1988).
 [8] Benner S., Nature 329:21-21(1987).
 20 [9] Kirby R., J. Mol. Evol. 30:489-492(1992).
 [10] Littler E., Stuart A.D., Chee M.S., Nature 358:160-162(1992).
 [11] Munoz-Dorado J., Inouye S., Inouye M., Cell 67:995-1006(1991).

25 835. (Asp_Glu_race)

Aspartate and glutamate racemases signatures

Cross-reference(s) PS00923; ASP_GLU_RACEMASE_1 PS00924;
 ASP_GLU_RACEMASE_2

30

Aspartate racemase (EC 5.1.1.13) and glutamate racemase (EC 5.1.1.3) are two evolutionary related bacterial enzymes that do not seem to require a cofactor for their activity [1].

Glutamate racemase, which interconverts L-glutamate into D-glutamate, is required for the

694

biosynthesis of peptidoglycan and some peptide-based antibiotics such as gramicidin S. In addition to characterized aspartate and glutamate racemases, this family also includes a hypothetical protein from *Erwinia carotovora* and one from *Escherichia coli* (ygeA). Two conserved cysteines are present in the sequence of these enzymes. They are expected to play a role in catalytic activity by acting as bases in proton abstraction from the substrate. Signature patterns were developed for both cysteines.

Consensus pattern: [IVA]-[LIVM]-x-C-x(0,1)-N-[ST]-[MSA]-[STH]-[LIVFYSTANK]

Consensus pattern: [LIVM](2)-x-[AG]-C-T-[DEH]-[LIVMFY]-[PNGRS]-x-[LIVM]

[1] Gallo K.A., Knowles J.R., Biochemistry 32:3981-3990(1993).

836. (ATP-sulfurylase)

ATP-sulfurylase

This family consists of ATP-sulfurylase or sulfate adenylyltransferase EC:2.7.7.4 some of which are part of a bifunctional polypeptide chain associated with adenosyl phosphosulphate (APS) kinase APS_kinase. Both enzymes are required for PAPS (phosphoadenosine-phosphosulfate) synthesis from inorganic sulphate [2]. ATP sulfurylase catalyses the synthesis of adenosine-phosphosulfate APS from ATP and inorganic sulphate [1].

Number of members: 37

[1] Kurima K, Warman ML, Krishnan S, Domowicz M, Krueger RC Jr, Deyrup A, Schwartz NB; Medline: 98337975 "A member of a family of sulfate-activating enzymes causes murine brachymorphism" [published erratum appears in Proc Natl Acad Sci U S A 1998 Sep 29;95(20):12071] Proc Natl Acad Sci U S A 1998;95:8681-8685.

[2] Rosenthal E, Leustek T; Medline: 96096529 "A multifunctional *Urechis caupo* protein, PAPS synthetase, has both ATP sulfurylase and APS kinase activities." Gene 1995;165:243-248.

837. (ATP-synt_F)

ATP synthase (F/14-kDa) subunit

5

This family includes 14-kDa subunit from vATPases [1], which is in the peripheral catalytic part of the complex [2]. The family also includes archaebacterial ATP synthase subunit F [3].

Number of members: 23

10

[1] Guo Y, Kaiser K, Wieczorek H, Dow JA; Medline: 96269411 "The *Drosophila melanogaster* gene *vha14* encoding a 14-kDa F-subunit of the vacuolar ATPase." *Gene* 1996;172:239-243.

[2] Peng SB, Crider BP, Tsai SJ, Xie XS, Stone DK; Medline: 96216416 "Identification of a 14-kDa subunit associated with the catalytic sector of clathrin-coated vesicle H⁺-ATPase." *J Biol Chem* 1996;271:3324-3327.

[3] Wilms R, Freiberg C, Wegerle E, Meier I, Mayer F, Muller V; Medline: 96324968 "Subunit structure and organization of the genes of the A1A0 ATPase from the Archaeon *Methanosarcina mazei* Go1." *J Biol Chem* 1996;271:18843-18852.

838. (CBD_4)

Starch binding domain

25 Number of members: 48

839. (CbiX)

30 The function of CbiX is uncertain, however it is found in cobalamin biosynthesis operons and so may have a related function. Some CbiX proteins contain a striking histidine-rich region at their C-terminus, which suggests that it might be involved in metal chelation [1].

Number of members: 6

[1] Raux E, Lanois A, Warren MJ, Rambach A, Thermes C; Medline: 98416126 "Cobalamin (vitamin B12) biosynthesis: identification and characterization of a *Bacillus megaterium* cobI operon." Biochem J 1998;335:159-166.

840. (Complex1_51K)

10 Respiratory-chain NADH dehydrogenase 51 Kd subunit signatures Cross-reference(s)
PS00644; COMPLEX1_51K_1 PS00645; COMPLEX1_51K_2

15 Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex there is one with a molecular weight of 51 Kd (in mammals), which is the second largest subunit of complex I and is a component of the iron-sulfur (IP) fragment of the enzyme. It seems to bind to NAD, FMN, and a 2Fe-2S cluster.

20 The 51 Kd subunit is highly similar to [3,4]:

- Subunit alpha of *Alcaligenes eutrophus* NAD-reducing hydrogenase (gene *hoxF*) which also binds to NAD, FMN, and a 2Fe-2S cluster.
- Subunit NQO1 of *Paracoccus denitrificans* NADH-ubiquinone oxidoreductase.
- 25 - Subunit F of *Escherichia coli* NADH-ubiquinone oxidoreductase (gene *nuoF*).

30 The 51 Kd subunit and the bacterial hydrogenase alpha subunit contains three regions of sequence similarities. The first one most probably corresponds to the NAD-binding site, the second to the FMN-binding site, and the third one, which contains three cysteines, to the iron-sulfur binding region. Signature patterns have been developed for the FMN-binding and for the 2Fe-2S binding regions.

Consensus pattern: G-[AM]-G-[AR]-Y-[LIVM]-C-G-[DE](2)-[STA](2)-[LIM](2)-[EN]- S

Consensus pattern: E-S-C-G-x-C-x-P-C-R-x-G [The three C's are putative 2Fe-2S ligands]

[1] Ragan C.I., Curr. Top. Bioenerg. 15:1-36(1987).

[2] Weiss H., Friedrich T., Hofhaus G., Preis D., Eur. J. Biochem. 197:563-576(1991).

5 [3] Fearnley I.M., Walker J.E. Biochim. Biophys. Acta 1140:105-134(1992).

[4] Weidner U., Geier S., Ptock A., Friedrich T., Leif H., Weiss H., J. Mol. Biol. 233:109-122(1993).

10 841. (DAP_epimerase)

Diaminopimelate epimerase signature

Cross-reference(s) PS01326; DAP_EPIMERASE

Diaminopimelate epimerase (EC 5.1.1.7) catalyzes the isomeriazation of L,L- to D,L-meso-diaminopimelate in the biosynthetic pathway leading from aspartate to lysine. This enzyme is a protein of about 30 Kd. Two conserved cysteines seem [1] to function as the acid and base in the catalytic mechanism. As a signature pattern, the region surrounding the first of these two active site cysteines were selected.

20 Consensus pattern: N-x-D-G-S-x(4)-C-G-N-[GA]-x-R [C is an active site residue] Sequences known to belong to this class detected by the pattern ALL, except for an Anabaena dapF which has a Ser instead of the active site Cys.

[1] Cirilli M., Zheng R., Scapin G., Blanchard J.S., Biochemistry 37:16452-16458(1998).

25

842. (DNA_gyraseB_C)

DNA topoisomerase II signature

30 Cross-reference(s) PS00177; TOPOISOMERASE_II

DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type II topoisomerases are ATP-dependent and act by passing a DNA segment through a transient double-strand break.

698

Topoisomerase II is found in phages, archaebacteria, prokaryotes, eukaryotes, and in African Swine Fever virus (ASF). In bacteriophage T4 topoisomerase II consists of three subunits (the product of genes 39, 52 and 60). In prokaryotes and in archaebacteria the enzyme, known as DNA gyrase, consists of two subunits (genes *gyrA* and *gyrB* [E2]). In some bacteria, a second type II topoisomerase has been identified; it is known as topoisomerase IV and is required for chromosome segregation, it also consists of two subunits (genes *parC* and *parE*). In eukaryotes, type II topoisomerase is a homodimer.

There are many regions of sequence homology between the different subtypes of topoisomerase II. The relation between the different subunits is shown in the following representation:

```

<-----About-1400-residues----->

[-----Protein 39-*-----][----Protein 52----]      Phage T4
[-----gyrB-----*-----][-----gyrA-----]  Prokaryote II
                                   Archaebacteria
[-----parE-----*-----][-----parD-----]  Prokaryote IV
[-----*-----] Eukaryote and
                                   ASF

```

*: Position of the pattern.

As a signature pattern for this family of proteins, a region that contains a highly conserved pentapeptide was selected. The pattern is located in *gyrB*, in *parE*, and in protein 39 of phage T4 topoisomerase.

Consensus pattern: [LIVMA]-x-E-G-[DN]-S-A-x-[STAG]

[1] Sternglanz R., Curr. Opin. Cell Biol. 1:533-535(1990).

[2] Bjornsti M.-A., Curr. Opin. Struct. Biol. 1:99-103(1991).

[3] Sharma A., Mondragon A., Curr. Opin. Struct. Biol. 5:39-47(1995).

[4] Roca J., Trends Biochem. Sci. 20:156-160(1995).

843. (DUF16)

Protein of unknown function

- 5 The function of this protein is unknown. It appears to only occur in *Mycoplasma pneumoniae*.

Number of members: 26

- 10 [1] Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R; Medline: 97105885
“Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*.”
Nucleic Acids Res 1996;24:4420-4449.

15 844. (DUF21)

Domain of unknown function

20 This transmembrane region has no known function. Many of the sequences in this family are annotated as hemolysins, however this is due to a similarity to Swiss:Q54318 that does not contain this domain. This domain is found in the N-terminus of the proteins adjacent to two intracellular CBS domains CBS.

Number of members: 42

25 845. (DUF56)

Integral membrane protein

30 The members of this family are putative integral membrane proteins. The function of the family is unknown, however the family includes Sec59 from yeast. Sec59 is a dolichol

700

kinase EC:2.7.1.108, but it is not clear if the enzymatic activity resides in this region or its N terminal region.

Number of members: 13

5

846. (DUF94)

Domain of unknown function

10

The function of this domain is unknown. It is found in both eukaryotes and archaeobacteria. The alignment contains a completely conserved aspartate residue that may be functionally important. The eukaryotic domains contains three conserved cysteines and a histidine that might be metal binding, however these are absent in the archaeobacterial proteins.

15
20

Number of members: 9

847. (FF)

FF domain

This domain may be involved in protein-protein interaction [1].

25 Number of members: 42

[1] Bedford MT, Leder P; Medline: 99322199 "The FF domain: a novel motif that often accompanies WW domains." Trends Biochem Sci 1999;24:264-265.

30

848. (FLO_LFY)

Floricaula / Leafy protein

701

This family consists of various plant development proteins which are homologues of floricaula (FLO) and Leafy (LFY) proteins which are floral meristem identity proteins. Mutations in the sequences of these proteins affect flower and leaf development.

5 Number of members: 16

[1] Hofer J, Turner L, Hellens R, Ambrose M, Matthews P, Michael A, Ellis N; Medline: 97411151 "UNIFOLIATA regulates leaf and flower morphogenesis in pea." Curr Biol 1997;7:581-587.

10 [2] Weigel D, Alvarez J, Smyth DR, Yanofsky MF, Meyerowitz EM; Medline: 92274452 "LEAFY controls floral meristem identity in Arabidopsis." Cell 1992;69:843-859.

849. (G-patch)

15 G-patch domain

This domain is found in a number of RNA binding proteins, and is also found in proteins that contain RNA binding domains. This suggests that this domain may have an RNA binding function. This domain has seven highly conserved glycines.

20

Number of members: 47

[1] Aravind L, Koonin EV; Medline: 10470032 "G-patch: a new conserved domain in eukaryotic RNA-processing proteins and type D retroviral polyproteins." Trends Biochem Sci 1999;24:342-344.

25

850. (Gram-ve_porins)

General diffusion Gram-negative porins signature

30

Cross-reference(s) PS00576; GRAM_NEG_PORIN

The outer membrane of Gram-negative bacteria acts as a molecular filter for hydrophilic compounds. Proteins, known as porins [1], are responsible for the 'molecular sieve' properties

of the outer membrane. Porins form large water- filled channels which allows the diffusion of hydrophilic molecules into the periplasmic space. Some porins form general diffusion channels that allows any solutes up to a certain size (that size is known as the exclusion limit) to cross the membrane, while other porins are specific for a solute and contain a binding site for that solute inside the pores (these are known as selective porins). As porins are the major outer membrane proteins, they also serve as receptor sites for the binding of phages and bacteriocins. General diffusion porins generally assemble as trimer in the membrane and the transmembrane core of these proteins is composed exclusively of beta strands [2]. It has been shown [3] that a number of general porins are evolutionary related, these porins are:

- Enterobacteria phoE.
- Enterobacteria ompC.
- Enterobacteria ompF.
- Enterobacteria nmpC.
- Bacteriophage PA-2 LC.
- Neisseria PI.A.
- Neisseria PI.B.

As a signature pattern a conserved region was selected, located in the C-terminal part of these proteins, which spans two putative transmembrane beta strands.

Consensus pattern: [LIVMFY]-x(2)-G-x(2)-Y-x-F-x-K-x(2)-[SN]-[STAV]-[LIVMFYW]- V

[1] Benz R., Bauer K., Eur. J. Biochem. 176:1-19(1988).

[2] Jap B.K., Walian P.J., Q. Rev. Biophys. 23:367-403(1990).

[3] Jeanteur D., Lakey J.H., Pattus F., Mol. Microbiol. 5:2153-2164(1991).

851. (HlyD)

HlyD family secretion proteins signature

Cross-reference(s) PS00543; HLYD_FAMILY

Gram-negative bacteria produce a number of proteins which are secreted into the growth medium by a mechanism that does not require a cleaved N-terminal signal sequence. These

proteins, while having different functions, require the help of two or more proteins for their secretion across the cell envelope. Amongst which a protein belonging to the ABC transporters family (see the relevant entry <PDOC00185>) and a protein belonging to a family which is currently composed [1 to 5] of the following members:

5	Gene	Species	Protein which is exported
	----	-----	-----
	hlyD	Escherichia coli	Hemolysin
	appD	A.pleuropneumoniae	Hemolysin
	lcnD	Lactococcus lactis	Lactococcin A
10	lktD	A.actinomycetemcomitans	Leukotoxin
		Pasteurella haemolytica	
	rtxD	A.pleuropneumoniae	Toxin-III
	cyaD	Bordetella pertussis	Calmodulin-sensitive adenylate cyclase-hemolysin (cyclolysin)
15	cvaA	Escherichia coli	Colicin V
	prtE	Erwinia chrysanthemi	Extracellular proteases B and C
	aprE	Pseudomonas aeruginosa	Alkaline protease
	emrA	Escherichia coli	Drugs and toxins
	yjcR	Escherichia coli	Unknown
20	These proteins are evolutionary related and consist of from 390 to 480 amino acid residues. They seem to be anchored in the inner membrane by a N-terminal transmembrane region. Their exact role in the secretion process is not yet known. The C-terminal section of these proteins is the best conserved region; a signature pattern from that region was derived.		
25	Consensus pattern: [LIVM]-x(2)-G-[LM]-x(3)-[STGAV]-x-[LIVMT]-x-[LIVMT]-[GE]-x-[KR]-x-[LIVMFYW](2)-x-[LIVMFYW](3)		
	Sequences known to belong to this class detected by the pattern ALL, except for emrA and yjcR.		
30	References:		
	[1] Gilson L., Mahanty H.K., Kolter R., EMBO J. 9:3875-3884(1990).		
	[2] Letoffe S., Delepelaire P., Wandersman C., EMBO J. 9:1375-1382(1990).		

[3] Stoddard G.W., Petzel J.P., van Belkum M.J., Kok J., McKay L.L., Appl. Environ. Microbiol. 58:1952-1961(1992).

[4] Duong F., Lazdunski A., Cami B., Murgier M., Gene 121:47-54(1992).

[5] Lewis K., Trends Biochem. Sci. 19:119-123(1994).

5

852. (IBR)

In Between Ring fingers

10 The IBR (In Between Ring fingers) domain is found to occur between pairs of ring fingers (zf-C3HC4). The function of this domain is unknown. This domain has also been called the C6HC domain and DRIL (for double RING finger linked) domain [2].

Number of members: 25

15 [1] Morett E, Bork P; Medline: 10366851 "A novel transactivation domain in parkin."Trends Biochem Sci 1999;24:229-231.

[2] van der Reijden BA, Erpelinck-Verschueren CA, Lowenberg B, Jansen JH; Medline: 99349709 "TRIADs: a new class of proteins with a novel cysteine-rich signature." Protein Sci 1999;8:1557-1561.

20

853. (IPPT)

IPP transferase

25 [1] Durand JM, Bjork GR, Kuwae A, Yoshikawa M, Sasakawa C; Medline: 97440126 "The modified nucleoside 2-methylthio-N6-isopentenyladenosine in tRNA of Shigella flexneri is required for expression of virulence genes." J Bacteriol 1997;179:5777-5782.

[2] Boguta M, Hunter LA, Shen WC, Gillman EC, Martin NC, Hopper AK; Medline: 94187700 "Subcellular locations of MOD5 proteins: mapping of sequences sufficient for targeting to mitochondria and demonstration that mitochondrial and nuclear isoforms commingle in the cytosol." Mol Cell Biol 1994;14:2298-2306.

30

[3] Gillman EC, Slusher LB, Martin NC, Hopper AK; Medline: 91203856 "MOD5 translation initiation sites determine N6-isopentenyladenosine modification of mitochondrial and cytoplasmic tRNA." Mol Cell Biol 1991;11:2382-2390.

5

854. (KE2)
KE2 family protein

10

The function of members of this family is unknown, although they have been suggested to contain a DNA binding leucine zipper motif [2].

Number of members: 9

15

[1] Ha H, Abe K, Artzt K; Medline: 92084131 "Primary structure of the embryo-expressed gene KE2 from the mouse H-2K region." Gene 1991;107:345-346.

[2] Shang HS, Wong SM, Tan HM, Wu M; Medline: 95129859 "YKE2, a yeast nuclear gene encoding a protein showing homology to mouse KE2 and containing a putative leucine-zipper motif." Gene 1994;151:197-201.

20

855. (Lipoprotein_6)
Prokaryotic membrane lipoprotein lipid attachment site

Cross-reference(s) PS00013; PROKAR_LIPOPROTEIN

25

In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):

30

- Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp).
- Escherichia coli lipoprotein-28 (gene nlpA).
- Escherichia coli lipoprotein-34 (gene nlpB).
- Escherichia coli lipoprotein nlpC.

- *Escherichia coli* lipoprotein nlpD.
- *Escherichia coli* osmotically inducible lipoprotein B (gene *osmB*).
- *Escherichia coli* osmotically inducible lipoprotein E (gene *osmE*).
- *Escherichia coli* peptidoglycan-associated lipoprotein (gene *pal*).
- 5 - *Escherichia coli* rare lipoproteins A and B (genes *rplA* and *rplB*).
- *Escherichia coli* copper homeostasis protein *cutF* (or *nlpE*).
- *Escherichia coli* plasmids *traT* proteins.
- *Escherichia coli* Col plasmids lysis proteins.
- A number of *Bacillus* beta-lactamases.
- 10 - *Bacillus subtilis* periplasmic oligopeptide-binding protein (gene *oppA*).
- *Borrelia burgdorferi* outer surface proteins A and B (genes *ospA* and *ospB*).
- *Borrelia hermsii* variable major protein 21 (gene *vmp21*) and 7 (gene *vmp7*).
- *Chlamydia trachomatis* outer membrane protein 3 (gene *omp3*).
- *Fibrobacter succinogenes* endoglucanase *cel-3*.
- 15 - *Haemophilus influenzae* proteins *Pal* and *Pcp*.
- *Klebsiella pullulunase* (gene *pulA*).
- *Klebsiella pullulunase* secretion protein *pulS*.
- *Mycoplasma hyorhinitis* protein *p37*.
- *Mycoplasma hyorhinitis* variant surface antigens A, B, and C (genes *vlpABC*).
- 20 - *Neisseria* outer membrane protein H.8.
- *Pseudomonas aeruginosa* lipopeptide (gene *lppL*).
- *Pseudomonas solanacearum* endoglucanase *egl*.
- *Rhodopseudomonas viridis* reaction center cytochrome subunit (gene *cytC*).
- *Rickettsia* 17 Kd antigen.
- 25 - *Shigella flexneri* invasion plasmid proteins *mxj* and *mxm*.
- *Streptococcus pneumoniae* oligopeptide transport protein A (gene *amiA*).
- *Treponema pallidum* 34 Kd antigen.
- *Treponema pallidum* membrane protein A (gene *tmpA*).
- *Vibrio harveyi* chitinase (gene *chb*).
- 30 - *Yersinia* virulence plasmid protein *yscJ*.
- Halocyanin from *Natrobacterium pharaonis* [4], a membrane associated copper-binding protein. This is the first archaebacterial protein known to be modified in such a fashion).

From the precursor sequences of all these proteins, a consensus pattern and a set of rules to identify this type of post-translational modification were derived.

Consensus pattern: {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1)

The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least one Lys or one Arg in the first seven positions of the sequence. Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT some 100 prokaryotic proteins. Some of them are not membrane lipoproteins, but at least half of them could be.

References

- [1] Hayashi S., Wu H.C., J. Bioenerg. Biomembr. 22:451-471(1990).
- [2] Klein P., Somorjai R.L., Lau P.C.K., Protein Eng. 2:15-20(1988).
- [3] von Heijne G., Protein Eng. 2:531-534(1989).
- [4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).

856. (Lipoprotein_7)

Adhesin lipoprotein

This family consists of the p50 and variable adherence-associated antigen (Vaa) adhesins from *Mycoplasma hominis*. *M. hominis* is a mycoplasma associated with human urogenital diseases, pneumonia, and septic arthritis [1]. An adhesin is a cell surface molecule that mediates adhesion to other cells or to the surrounding surface or substrate. The Vaa antigen is a 50-kDa surface lipoprotein that has four tandem repetitive DNA sequences encoding a periodic peptide structure, and is highly immunogenic in the human host [1]. p50 is also a 50-kDa lipoprotein, having three repeats A,B and C, that may be a tetramer of 191-kDa in its native environment [2].

Number of members: 18

[1] Zhang Q, Wise KS; Medline: 96294788 “Molecular basis of size and antigenic variation of a *Mycoplasma hominis* adhesin encoded by divergent *vaa* genes. “ *Infect Immun* 1996;64:2737-2744.

[2] Henrich B, Kitzerow A, Feldmann RC, Schaal H, Hadding U; Medline: 97047675 “Repetitive elements of the *Mycoplasma hominis* adhesin p50 can be differentiated by monoclonal antibodies.” *Infect Immun* 1996;64:4027-4034.

857. (MaoC_like)
MaoC like domain

The MaoC protein is found to share similarity with a wide variety of enzymes; estradiol 17 beta-dehydrogenase 4, peroxisomal hydratase-dehydrogenase-epimerase, fatty acid synthase beta subunit. All these enzymes contain other domains. This domain is also present in the NodN nodulation protein N. No specific function has been assigned to this region of any of these proteins. The *maoC* gene is part of a operon with *maoA* which is involved in the synthesis of monoamine oxidase [1].

Number of members: 46

[1] Sugino H, Sasaki M, Azakami H, Yamashita M, Murooka Y Medline: 96235221 “A monoamine-regulated *Klebsiella aerogenes* operon containing the monoamine oxidase structural gene (*maoA*) and the *maoC* gene.” *J Bacteriol* 1992;174:2485-2492.

858. (MSP)
Manganese-stabilizing protein / photosystem II polypeptide

This family consists of the 33 KDa photosystem II polypeptide from the oxygen evolving complex (OEC) of plants and cyanobacteria. The protein is also known as the manganese-stabilizing protein as it is associated with the manganese complex of the OEC and may provide the ligands for the complex [1].

Number of members: 17

[1] Philbrick JB, Zilinskas BA; Medline: 88334494 "Cloning, nucleotide sequence and
mutational analysis of the gene encoding the Photosystem II manganese-stabilizing
polypeptide of *Synechocystis* 6803." *Mol Gen Genet* 1988;212:418-425.

859. (NAC)

[1] Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV;
Medline: 99342100 "Comparative genomics of the Archaea (Euryarchaeota): evolution of
conserved protein families, the stable core, and the variable shell." *Genome Res* 1999;9:608-
628.

Number of members: 27

860. (Nop)

Putative snoRNA binding domain

This family consists of various Pre RNA processing ribonucleoproteins. The function of the
aligned region is unknown however it may be a common RNA or snoRNA or Nop1p binding
domain. Nop5p (Nop58p) Swiss:Q12499 from yeast is the protein component of a
ribonucleoprotein protein required for pre-18s rRNA processing and is suggested to function
with Nop1p in a snoRNA complex [1]. Nop56p Swiss:O00567 and Nop5p interact with
Nop1p and are required for ribosome biogenesis [2]. Prp31p Swiss:p49704 is required for
pre-mRNA splicing in *S. cerevisiae* [3].

Number of members: 23

[1] Wu P, Brockenbrough JS, Metcalfe AC, Chen S, Aris JP; Medline: 98298165 "Nop5p is a small nucleolar ribonucleoprotein component required for pre- 18 S rRNA processing in yeast." J Biol Chem 1998;273:16453-16463.

[2] Gautier T, Berges T, Tollervy D, Hurt E; Medline: 8038777 "Nucleolar KKE/D repeat proteins Nop56p and Nop58p interact with Nop1p and are required for ribosome biogenesis." Mol Cell Biol 1997;17:7088-7098.

[3] Weidenhammer EM, Singh M, Ruiz-Noriega M, Woolford JL Jr; Medline: 96184869 "The PRP31 gene encodes a novel protein required for pre-mRNA splicing in *Saccharomyces cerevisiae*." Nucleic Acids Res 1996;24:1164-1170.

861. (Nramp)

Natural resistance-associated macrophage protein

The natural resistance-associated macrophage protein (NRAMP) family consists of Nramp1, Nramp2, and yeast proteins Smf1 and Smf2. The NRAMP family is a novel family of functional related proteins defined by a conserved hydrophobic core of ten transmembrane domains [5]. This family of membrane proteins are divalent cation transporters. Nramp1 is an integral membrane protein expressed exclusively in cells of the immune system and is recruited to the membrane of a phagosome upon phagocytosis [1]. By controlling divalent cation concentrations Nramp1 may regulate the interphagosomal replication of bacteria [1]. Mutations in Nramp1 may genetically predispose an individual to susceptibility to diseases including leprosy and tuberculosis conversely this might however provide protection from rheumatoid arthritis [1]. Nramp2 is a multiple divalent cation transporter for Fe²⁺, Mn²⁺ and Zn²⁺ amongst others it is expressed at high levels in the intestine; and is major transferrin-independent iron uptake system in mammals [1]. The yeast proteins Smf1 and Smf2 may also transport divalent cations [3].

Number of members: 36

[1] Govoni G, Gros P; Medline: 98383996 "Macrophage NRAMP1 and its role in resistance to microbial infections." Inflamm Res 1998;47:277-284.

711

[2] Agranoff DD, Krishna S Medline: 98294035 "Metal ion homeostasis and intracellular parasitism." Mol Microbiol 1998;28:403-412.

[3] Pinner E, Gruenheid S, Raymond M, Gros P; Medline: 98030569 "Functional complementation of the yeast divalent cation transporter family SMF by NRAMP2, a member of the mammalian natural resistance- associated macrophage protein family." J Biol Chem 1997;272:28933-28938.

[4] Cellier M, Belouchi A, Gros P; Medline: 96402487 "Resistance to intracellular infections: comparative genomic analysis of Nramp." Trends Genet 1996;12:201-204.

[5] Cellier M, Prive G, Belouchi A, Kwan T, Rodrigues V, Chia W, Gros P; Medline: 96036029 "Nramp defines a family of membrane proteins." Proc Natl Acad Sci U S A 1995;92:10089-10093.

862. (NTP_transf_2)

Nucleotidyltransferase domain

Members of this family belong to a large family of nucleotidyltransferases [1].

Number of members: 83

[1] Holm L, Sander C; Medline: 96005605 "DNA polymerase beta belongs to an ancient nucleotidyltransferase superfamily." Trends Biochem Sci 1995;20:345-347.

863. (Paramyxo_P)

Paramyxovirus P phosphoprotein

This family consists of paramyxovirus P phosphoprotein from sendai virus and human and bovine parainfluenza viruses. The P protein is an essential part of the viral RNA polymerase complex formed from the P and L proteins [1]. The exact role of the P protein in this complex is unknown but it is involved in multiple protein-protein interactions and binding the polymerase complex to the nucleocapsid or ribonucleoprotein template [1]. It also appears to

712

be important for the proper folding of the L protein [1]. The paramyxoviruses have a negative sense ssRNA genome [1].

Number of members: 15

5

[1] Bowman MC, Smallwood S, Moyer SA; Medline: 99329169 "Dissection of Individual Functions of the Sendai Virus Phosphoprotein in Transcription." J Virol 1999;73:6474-6483.

[2] Matsuoka Y, Curran J, Pelet T, Kolakofsky D, Ray R, Compans RW; Medline: 91237868 "The P gene of human parainfluenza virus type 1 encodes P and C proteins but not a cysteine-rich V protein." J Virol 1991;65:3406-3410.

10

864. (Patatin)

This family consists of various patatin glycoproteins from plants. The patatin protein accounts for up to 40% of the total soluble protein in potato tubers [2]. Patatin is a storage protein but it also has the enzymatic activity of lipid acyl hydrolase, catalysing the cleavage of fatty acids from membrane lipids [2].

Number of members: 21

[1] Banfalvi Z, Kostyal Z, Barta E; Medline: 95107249 "Solanum brevidens possesses a non-sucrose-inducible patatin gene." Mol Gen Genet 1994;245:517-522.

[2] Mignery GA, Pikaard CS, Park WD; Medline: 88226014 "Molecular characterization of the patatin multigene family of potato." Gene 1988;62:27-44.

865. (Pentapeptide_2)

Pentapeptide repeats (8 copies)

These repeats are found in many mycobacterial proteins. These repeats are most common in the PPE family of proteins, where they are found in the MPTR subfamily of PPE proteins. The function of these repeats is unknown. The repeat can be approximately described as

30

713

XNXGX, where X can be any amino acid. These repeats are similar to Pentapeptide [1], however it is not clear if these two families are structurally related.

Number of members: 362

5

[1] Bateman A, Murzin A, Teichmann SA; Medline: 98318059 "Structure and distribution of pentapeptide repeats in bacteria." Protein Sci 1998;7:1477-1480.

[2] Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Barrell BG; Medline: 98295987 "Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence." Nature 1998;393:537-544.

10

15

866. (Peptidase_C13)

Peptidase C13 family

This family of peptidases is known as the hemoglobinase family because it contains a globin degrading enzyme from blood parasites Swiss:P42665. However relatives are found in plants and other organisms that have other functions. Members of this family are asparaginyl peptidases [1].

20

Number of members: 26

25

[1] Chen JM, Dando PM, Rawlings ND, Brown MA, Young NE, Stevens RA, Hewitt E, Watts C, Barrett AJ; Medline: 97218252 "Cloning, isolation, and characterization of mammalian legumain, an asparaginyl endopeptidase." J Biol Chem 1997;272:8090-8098.

30

867. (Pro_dh)

Proline dehydrogenase

Number of members: 25

[1] Ling M, Allen SW, Wood JM; Medline: 95055736 "Sequence analysis identifies the proline dehydrogenase and delta 1- pyrroline-5-carboxylate dehydrogenase domains of the multifunctional Escherichia coli PutA protein." J Mol Biol 1994;243:950-956.

868. (PsbP)

This family consists of the 23 kDa subunit of oxygen evolving system of photosystem II or PsbP from various plants (where it is encoded by the nuclear genome) and Cyanobacteria. The 23 KDa PsbP protein is required for PSII to be fully operational in vivo, it increases the affinity of the water oxidation site for Cl- and provides the conditions required for high affinity binding of Ca²⁺ [2].

Number of members: 25

[1] Rova EM, Mc Ewen B, Fredriksson PO, Styring S; Medline: 97067138 "Photoactivation and photoinhibition are competing in a mutant of Chlamydomonas reinhardtii lacking the 23-kDa extrinsic subunit of photosystem II." J Biol Chem 1996;271:28918-28924.

[2] Kochhar A, Khurana JP, Tyagi AK; Medline: 97191538 "Nucleotide sequence of the psbP gene encoding precursor of 23-kDa polypeptide of oxygen-evolving complex in Arabidopsis thaliana and its expression in the wild-type and a constitutively photomorphogenic mutant." DNA Res 1996;3:277-285.

869. (PUA)

The PUA domain named after PseudoUridine synthase and Archaeosine transglycosylase, was detected in archaeal and eukaryotic pseudouridine synthases, archaeal archaeosine synthases, a family of predicted ATPases that may be involved in RNA modification, a family of predicted archaeal and bacterial rRNA methylases. Additionally, the PUA domain was detected in a family of eukaryotic proteins that also contain a domain homologous to the

715

translation initiation factor eIF1/SUI1; these proteins may comprise a novel type of translation factors. Unexpectedly, the PUA domain was detected also in bacterial and yeast glutamate kinases; this is compatible with the demonstrated role of these enzymes in the regulation of the expression of other genes [1]. It is predicted that the PUA domain is an RNA binding domain.

Number of members: 48

[1] Aravind L, Koonin EV; Medline: 99193178 "Novel predicted RNA-binding domains associated with the translation machinery." J Mol Evol 1999;48:291-302.

870. (RF1)

eRF1-like proteins

Members of this family are peptide chain release factors. The eukaryotic Release Factor 1 proteins (eRF1s) are involved in termination of translation. The eRF1 protein is functional for all stop codons and appears to abolish read-through of these codons. This family also includes other proteins for which the precise molecular function is unknown. Many of them are from Archaeobacteria. These proteins may also be involved in translation termination but this awaits experimental verification. Number of members: 25

[1] Frolova L, Le Goff X, Rasmussen HH, Cheperegin S, Drugeon G, Kress M, Arman I, Haenni AL, Celis JE, Philippe M, et al; Medline: 95082951 "A highly conserved eukaryotic protein family possessing properties of polypeptide chain release factor" [see comments] Nature 1994;372:701-703.

[2] Drugeon G, Jean-Jean O, Frolova L, Le Goff X, Philippe M, Kisselev L, Haenni AL; Medline: 97315314 "Eukaryotic release factor 1 (eRF1) abolishes readthrough and competes with suppressor tRNAs at all three termination codons in messenger RNA." Nucleic Acids Res 1997;25:2254-2258.

871. (Ribosomal_L14e)Ribosomal protein L14

This family includes the eukaryotic ribosomal protein L14.

Number of members: 15

5 872. (Ribosomal_S27)

Ribosomal protein S27a

This family of ribosomal proteins consists mainly of the 40S ribosomal protein S27a which is synthesized as a C-terminal extension of ubiquitin (CEP). The S27a domain compromises the C-terminal half of the protein. The synthesis of ribosomal proteins as extensions of ubiquitin promotes their incorporation into nascent ribosomes by a transient metabolic stabilization and is required for efficient ribosome biogenesis [3]. The ribosomal extension protein S27a contains a basic region that is proposed to form a zinc finger; its fusion gene is proposed as a mechanism to maintain a fixed ratio between ubiquitin necessary for degrading proteins and ribosomes a source of proteins [2].

Number of members: 36

20 873. (Spermine_synth)

Spermine/spermidine synthase

Spermine and spermidine are polyamines. This family includes spermidine synthase that catalyses the fifth (last) step in the biosynthesis of spermidine from arginine, and spermine synthase.

Number of members: 39

[1] Mezquita J, Pau M, Mezquita C; Medline: 97449308 "Characterization and expression of two chicken cDNAs encoding ubiquitin fused to ribosomal proteins of 52 and 80 amino acids." Gene 1997;195:313-319.

[2] Redman KL, Rechsteiner M; Medline: 89181932 "Identification of the long ubiquitin extension as ribosomal protein S27a." Nature 1989;338:438-440.

[3] Finley D, Bartel B, Varshavsky A; Medline: 89181925 "The tails of ubiquitin precursors are ribosomal proteins whose fusion to ubiquitin facilitates ribosome biogenesis." Nature 1989;338:394-401.

5

874. (Surp)

Surp module

10

[1] Denhez F, Lafyatis R; Medline: 94266805 "Conservation of regulated alternative splicing and identification of functional domains in vertebrate homologs to the Drosophila splicing regulator, suppressor-of-white-apricot." J Biol Chem 1994;269:16170-16179.

15

This domain is also known as the SWAP domain. SWAP stands for Suppressor-of-White-APricot. It has been suggested that these domains may be RNA binding [1].

Number of members: 32

20

875. (TFIIE)

TFIIE alpha subunit

25

The general transcription factor TFIIE has an essential role in eukaryotic transcription initiation together with RNA polymerase II and other general factors. Human TFIIE consists of two subunits TFIIE-alpha Swiss:P29083 and TFIIE-beta Swiss:P29084 and joins the preinitiation complex after RNA polymerase II and TFIIF [1]. This family consists of the conserved amino terminal region of eukaryotic TFIIE-alpha [2] and proteins from archaeobacteria that are presumed to be TFIIE-alpha subunits also Swiss:O29501 [3].

30

Number of members: 12

[1] Ohkuma Y, Sumimoto H, Hoffmann A, Shimasaki S, Horikoshi M, Roeder RG; Medline: 92065982 "Structural motifs and potential sigma homologies in the large subunit of human general transcription factor TFIIE." Nature 1991;354:398-401.

[2] Ohkuma Y, Hashimoto S, Roeder RG, Horikoshi M; Medline: 93087200 Identification of two large subdomains in TFIIE-alpha on the basis of homology between *Xenopus* and human sequences. *Nucleic Acids Res* 1992;20:5838-5838.

[3] Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC, Fleischmann RD, Quackenbush J, Lee NH, Sutton GG, Gill S, Kirkness EF, Dougherty BA, McKenney K, Adams MD, Loftus B, Venter JC, et al; Medline: 98049343 "The complete genome sequence of the hyperthermophilic, sulphate- reducing archaeon *Archaeoglobus fulgidus*." *Nature* 1997;390:364-370.

876. (Transglut_core)

Cross-reference(s) PS00547; TRANSGLUTAMINASES

Transglutaminases (EC 2.3.2.13) (TGase) [1,2] are calcium-dependent enzymes that catalyze the cross-linking of proteins by promoting the formation of isopeptide bonds between the gamma-carboxyl group of a glutamine in one polypeptide chain and the epsilon-amino group of a lysine in a second polypeptide chain. TGases also catalyze the conjugation of polyamines to proteins. The best known transglutaminase is blood coagulation factor XIII, a plasma tetrameric protein composed of two catalytic A subunits and two non-catalytic B subunits.

Factor XIII is responsible for cross-linking fibrin chains, thus stabilizing the fibrin clot. Other forms of transglutaminases are widely distributed in various organs, tissues and body fluids.

Sequence data is available for the following forms of TGase:

- Transglutaminase K (Tgase K), a membrane-bound enzyme found in mammalian epidermis and important for the formation of the cornified cell envelope (gene TGM1).
- Tissue transglutaminase (TGase C), a monomeric ubiquitous enzyme located in the cytoplasm (gene TGM2).
- Transglutaminase 3, responsible for the later stages of cell envelope formation in the epidermis and the hair follicle (gene TGM3).
- Transglutaminase 4 (gene TGM4).

A conserved cysteine is known to be involved in the catalytic mechanism of TGases. The erythrocyte membrane band 4.2 protein, which probably plays an important role in regulating the shape of erythrocytes and their mechanical properties, is evolutionary related to TGases. However the active site cysteine is substituted by an alanine and the 4.2 protein does not show TGase activity.

Consensus pattern:[GT]-Q-[CA]-W-V-x-[SA]-[GA]-[IVT]-x(2)-T-x-[LMSC]-R-[CSA]-[LV]-G [The first C is the active site residue] Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROT NONE.

- [1] Ichinose A., Bottenus R.E., Davie E.W. J. Biol. Chem. 265:13411-13414(1990).
- [2] Greenberg C.S., Birckbichler P.J., Rice R.H. FASEB J. 5:3071-3077(1991).

877. (TruB_N)

TruB family pseudouridylate synthase (N terminal domain)

Members of this family are involved in modifying bases in RNA molecules. They carry out the conversion of uracil bases to pseudouridine. This family includes TruB, a pseudouridylate synthase that specifically converts uracil 55 to pseudouridine in most tRNAs. This family also includes Cbf5p that modifies rRNA [2].

Number of members: 33

- [1] Nurse K, Wrzesinski J, Bakin A, Lane BG, Ofengand J; Medline: 96079944 "Purification, cloning, and properties of the tRNA psi 55 synthase from Escherichia coli." RNA 1995;1:102-112.
- [2] Lafontaine DLJ, Bousquet-Antonelli C, Henry Y, Caizergues-Ferrer M, Tollervey D; Medline: 98139521 "The box H + ACA snoRNAs carry Cbf5p, the putative rRNA pseudouridine synthase." Genes Dev 1998;12:527-537.

878. (UDPGP)

UTP--glucose-1-phosphate uridylyltransferase

This family consists of UTP--glucose-1-phosphate uridylyltransferases, EC:2.7.7.9. Also known as UDP-glucose pyrophosphorylase (UDPGP) and Glucose-1-phosphate uridylyltransferase. UTP--glucose-1-phosphate uridylyltransferase catalyses the interconversion of MgUTP + glucose-1-phosphate and UDP-glucose + MgPPi [1]. UDP-glucose is an important intermediate in mammalian carbohydrate interconversion involved in various metabolic roles depending on tissue type [1]. In Dictyostelium (slime mold) mutants in this enzyme abort the development cycle [2]. Also within the family is UDP-N-acetylglucosamine Swiss:Q16222 or AGX1 [3] and two hypothetical proteins from Borrelia burgdorferi the lyme disease spirochaete Swiss:O51893 and Swiss:O51036.

Number of members: 18

[1] Duggleby RG, Chao YC, Huang JG, Peng HL, Chang HY; Medline: 96202932 "Sequence differences between human muscle and liver cDNAs for UDPglucose pyrophosphorylase and kinetic properties of the recombinant enzymes expressed in Escherichia coli." Eur J Biochem 1996;235:173-179.

[2] Ragheb JA, Dottin RP; Medline: 87231075 "Structure and sequence of a UDP glucose pyrophosphorylase gene of Dictyostelium discoideum." Nucleic Acids Res 1987;15:3891-3906.

[3] Mio T, Yabe T, Arisawa M, Yamada-Okabe H; Medline: 98269105 "The eukaryotic UDP-N-acetylglucosamine pyrophosphorylases. Gene cloning, protein expression, and catalytic mechanism. J Biol Chem 1998;273:14392-14397.

879. (UPF004)

Uncharacterized protein family UPF0044 signature

Cross-reference(s) PS01301; UPF0044

The following uncharacterized proteins have been shown [1] to be highly similar:

- Bacillus subtilis hypothetical protein yqeI.

721

- Escherichia coli hypothetical protein yhbY and HI1333, the corresponding Haemophilus influenzae protein.

- Methanococcus jannaschii hypothetical protein MJ0652.

These are small proteins of 10 to 15 Kd. They can be picked up in the database

5 by the following pattern. This pattern is located in the N-terminal part of these proteins.

Consensus pattern: L-[ST]-x(3)-K-x(3)-[KR]-[SGA]-x-[GA]-H-x-L-x-P-[LIV]-x(2)-[LIV]-
[GA]-x(2)-G Sequences known to belong to this class detected by the patternALL. Other
10 sequence(s) detected in SWISS-PROT NONE.

880. (zf-A20)

A20-like zinc finger

15 A20- (an inhibitor of cell death)-like zinc fingers. The zinc finger mediates self-association in A20. These fingers also mediate IL-1-induced NF-kappa B activation.

Number of members: 22

[1] Heyninck K, Beyaert R; Medline: 99126071 "The cytokine-inducible zinc finger protein A20 inhibits IL-1-induced NF- kappaB activation at the level of TRAF6. FEBS Lett 1999;442:147-150.

[2] De Valck D, Heyninck K, Van Crielinge W, Contreras R, Beyaert R, Fiers W; Medline: 96390831 "A20, an inhibitor of cell death, self-associates by its zinc finger domain." FEBS Lett 1996;384:61-64.

[3] Song HY, Rothe M, Goeddel DV; Medline: 96270609 "The tumor necrosis factor-inducible zinc finger protein A20 interacts with TRAF1/TRAF2 and inhibits NF-kappaB activation. Proc Natl Acad Sci U S A 1996;93:6721-6725.

30 [4] Opipari AW Jr, Boguski MS, Dixit VM; Medline: 90368626 "The A20 cDNA induced by tumor necrosis factor alpha encodes a novel type of zinc finger protein." J Biol Chem 1990;265:14705-14708.

881. (zf-PARP)

Poly(ADP-ribose) polymerase zinc finger domain

5 Cross-reference(s) PS00347; PARP_ZN_FINGER_1 PS50064; PARP_ZN_FINGER_2

Poly(ADP-ribose) polymerase (EC 2.4.2.30) (PARP) [1,2] is a eukaryotic enzyme that catalyzes the covalent attachment of ADP-ribose units from NAD(+) to various nuclear acceptor proteins. This post-translational modification of nuclear proteins is dependent
10 on DNA. It appears to be involved in the regulation of various important cellular processes such as differentiation, proliferation and tumor transformation as well as in the regulation of the molecular events involved in the recovery of the cell from DNA damage. Structurally, PARP, about 1000 amino-acids residues long, consists of three distinct domains: an N-terminal zinc-dependent DNA-binding domain, a central automodification
15 domain and a C-terminal NAD-binding domain. The DNA-binding region contains a pair of zinc finger domains which have been shown to bind DNA in a zinc-dependent manner. The zinc finger domains of PARP seem to bind specifically to single-stranded DNA. DNA ligase III [3] contains, in its N-terminal section, a single copy of a zinc finger highly similar to those of PARP.

Consensus pattern: C-[KR]-x-C-x(3)-I-x-K-x(3)-[RG]-x(16,18)-W-[FYH]-H-x(2)-C [The
20 three C's and the H are zinc ligands] Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROT NONE. Sequences known to belong to this class detected by the profile ALL. Other sequence(s) detected in SWISS-
25 PROTNONE.

Note: This documentation entry is linked to both signature patterns and a profile. As the profile is much more sensitive than the patterns, you should use it if you have access to the necessary software tools to do so.

[1] Althaus F.R., Richter C.R. Mol. Biol. Biochem. Biophys. 37:1-126(1987).

[2] de Murcia G., Menissier de Murcia J. Trends Biochem. Sci. 19:172-176(1994).

[3] Wei Y.-F., Robins P., Carter K., Caldecott K., Pappin D.J.C., Yu G.-L., Wang R.-P., Shell B.K., Nash R.A., Schar P., Barnes D.E., Haseltine W.A., Lindahl T. Mol. Cell. Biol. 15:3206-3216(1995).

5 882. Adenylylsulfate kinase (APS_kinase)

Enzyme that catalyses the phosphorylation of adenylylsulfate to 3'-phosphoadenylylsulfate. This domain contains an ATP binding P-loop motif. Number of members: 34

10 [1] MacRae IJ, Rose AB, Segel IH; Medline: 99003196 "Adenosine 5'-phosphosulfate kinase from *Penicillium chrysogenum*. site- directed mutagenesis at putative phosphoryl-accepting and ATP P-loop residues. J Biol Chem 1998;273:28583-28589.

883. DNA polymerase family B signature DNA_POLYMERASE_B (DNA_pol_B)

Replicative DNA polymerases (EC 2.7.7.7) are the key enzymes catalyzing the accurate replication of DNA. They require either a small RNA molecule or a protein as a primer for the de novo synthesis of a DNA chain. On the basis of sequence similarity, a number of DNA polymerases have been grouped [1 to 7] under the designation of DNA polymerase family B. These are:

- Higher eukaryotes polymerases alpha.
- Higher eukaryotes polymerases delta.
- Yeast polymerase I/alpha (gene POL1), polymerase II/epsilon (gene POL2), polymerase III/delta (gene POL3) and polymerase REV3.
- *Escherichia coli* polymerase II (gene *dinA* or *polB*).
- Archaeobacterial polymerases.
- Polymerases of viruses from the herpesviridae family.
- Polymerases from Adenoviruses.
- Polymerases from Baculoviruses.
- Polymerases from Chlorella viruses.
- Polymerases from Poxviruses.
- Bacteriophage T4 polymerase.
- Podoviridae bacteriophages Phi-29, M2 and PZA polymerase.
- Tectiviridae bacteriophage PRD1 polymerase.

- Polymerases encoded on mitochondrial linear DNA plasmids in various fungi and plants (Kluyveromyces lactis pGKL1 and pGKL2, Agaricus bitorquis pEM, Ascobolus immersus pAI2, Claviceps purpurea pCLK1, Neurospora Kalilo and Maranhar, maize S-1, etc).

- 5 Six regions of similarity (numbered from I to VI) are found in all or a subset of the above polymerases. The most conserved region (I) includes a conserved tetrapeptide with two aspartate residues. Its function is not yet known. However, it has been suggested [3] that it may be involved in binding a magnesium ion. This conserved region was selected as a signature for this family of DNA polymerases.

10

Consensus pattern [YA]-[GLIVMSTAC]-D-T-D-[SG]-[LIVMFTC]-x-[LIVMSTAC]
Sequences known to belong to this class detected by the patternALL, except for yeast polymerase II/epsilon, Agaricus bitorquis pEM and Sulfolobus solfataricus polymerase II.

15

[1] Jung G., Leavitt M.C., Hsieh J.-C., Ito J. Proc. Natl. Acad. Sci. U.S.A. 84:8287-8291(1987).

[2] Bernad A., Zaballos A., Salas M., Blanco L. EMBO J. 6:4219-4225(1987).

[3] Argos P. Nucleic Acids Res. 16:9909-9916(1988).

[4] Wang T.S.-F., Wong S.W., Korn D. FASEB J. 3:14-21(1989).

20

[5] Delarue M., Poch O., Todro N., Moras D., Argos P. Protein Eng. 3:461-467(1990).

[6] Ito J., Braithwaite D.K. Nucleic Acids Res. 19:4045-4057(1991).

[7] Braithwaite D.K., Ito J. Nucleic Acids Res. 21:787-802(1993).

25

884. DNA polymerase family X signature - DNA_POLYMERASE_X (DNA_polymeraseX)

DNA polymerases (EC 2.7.7.7) can be classified, on the basis of sequence similarity [1], into at least four different groups: A, B, C and X. DNA polymerases that belong to family X are listed below [2]:

30

- Vertebrate polymerase beta, involved in DNA repair.

- Yeast polymerase IV (POL4) [3], an enzyme with similar characteristics to that of the mammalian polymerase beta.

725

- Terminal deoxynucleotidyltransferase (TdT) (EC 2.7.7.31). TdT catalyzes the elongation of polydeoxynucleotide chains by terminal addition. One of the functions of this enzyme is the addition of nucleotides at the junction of rearranged Ig heavy chain and T cell receptor gene segments during the maturation of B and T cells.

- 5 - African Swine Fever virus protein O174L [4].
- Fission yeast hypothetical protein SpAC2F7.06c.

These enzymes are small (about 40 Kd) compared with other polymerases and their reaction mechanism operates via a distributive mode, i.e. they dissociate from the template-primer after addition of each nucleotide.

As a signature pattern for this family of DNA polymerases, a highly conserved region that contains a conserved arginine and two conserved aspartic acid residues were selected. The latter together with the arginine have been shown [5] to be involved in primer binding in polymerase beta.

Consensus pattern G-[SG]-[LFY]-x-R-[GE]-x(3)-[SGCL]-x-D-[LIVM]-D- [LIVMFY](3)-x(2)-[SAP] Sequences known to belong to this class detected by the patternALL.

- [1] Ito J., Braithwaite D.K. Nucleic Acids Res. 19:4045-4057(1991).
- [2] Matsukage A., Nishikawa K., Ooi T., Seto Y., Yamaguchi M. J. Biol. Chem. 262:8960-8962(1987).
- [3] Prasad R., Widen S.G., Singhal R.K., Watkins J., Prakash L., Wilson S.H. Nucleic Acids Res. 21:5301-5307(1993).
- [4] Yanez R.J., Rodriguez J.M., Nogal M.L., Yuste L., Enriquez C., Rodriguez J.F., Vinuela E. Virology 208:249-278(1995).
- [5] Date T., Yamamoto S., Tanihara K., Nishimoto Y., Matsukage A. Biochemistry 30:5286-5292(1991).

885. DUF14 - Domain of unknown function

This domain is found in glutamate synthase, tungsten formylmethanofuran dehydrogenase subunit c (FwdC) and molybdenum formylmethanofuran dehydrogenase subunit c (FmdC). It has no known function. Number of members: 52

[1] Hochheimer A, Hedderich R, Thauer RK; Medline: 99035764. "The formylmethanofuran dehydrogenase isoenzymes in *Methanobacterium wolfei* and *Methanobacterium thermoautotrophicum*: induction of the molybdenum isoenzyme by molybdate and constitutive synthesis of the tungsten isoenzyme." Arch Microbiol 1998;170:389-393.

886. DUF18-Domain of unknown function

This domain of unknown function is found in several *C. elegans* proteins. The domain is 120 amino acids long and rich in cysteine residues. There are 16 conserved cysteine positions in the domain. Number of members: 34

887. DUF27-Domain of unknown function

This domain is found in a number of otherwise unrelated proteins. This domain is found at the C-terminus of the macro-H2A histone protein Swiss:Q02874. This domain is found in the non-structural proteins of several types of ssRNA viruses such as NSP2 from alphaviruses Swiss:P03317. This domain is also found on its own in a family of proteins from bacteria Swiss:P75918, archaeobacteria Swiss:O59182 and eukaryotes Swiss:Q17432, suggesting that it is involved in an important and ubiquitous cellular process. Number of members: 66

888. DUF37-Domain of unknown function

This domain is found in short (70 amino acid) hypothetical proteins from various bacteria. The domain contains three conserved cysteine residues. Swiss:Q44066 from *Aeromonas hydrophila* has been found to have hemolytic activity (unpublished). Number of members: 19

889. EGF-like domain signatures. (EGF-like)

A sequence of about thirty to forty amino-acid residues long found in the sequence of epidermal growth factor (EGF) has been shown [1 to 6] to be present, in a more or less conserved form, in a large number of other, mostly animal proteins. The proteins currently known to contain one or more copies of an EGF-like pattern are listed below.

- Adipocyte differentiation inhibitor (gene PREF-1) from mouse (6 copies).
- Agrin, a basal lamina protein that causes the aggregation of acetylcholine receptors on cultured muscle fibers (4 copies).

- Amphiregulin, a growth factor (1 copy).
- Betacellulin, a growth factor (1 copy).
- Blastula proteins BP10 and Span from sea urchin which are thought to be involved in pattern formation (1 copy).
- 5 - BM86, a glycoprotein antigen of cattle tick (7 copies).
- Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation and which expresses metalloendopeptidase activity (1-2 copies). Homologous proteins are found in sea urchin - suBMP (1 copy) - and in Drosophila - the dorsal-ventral patterning protein tolloid (2 copies).
- 10 - Caenorhabditis elegans developmental proteins lin-12 (13 copies) and glp-1 (10 copies).
- Caenorhabditis elegans APX-1 protein, a patterning protein (4.5 copies).
- Calcium-dependent serine proteinase (CASP) which degrades the extracellular matrix proteins type I and IV collagen and fibronectin (1 copy).
- Cartilage matrix protein CMP (1 copy).
- 15 - Cartilage oligomeric matrix protein COMP (4 copies).
- Cell surface antigen 114/A10 (3 copies).
- Cell surface glycoprotein complex transmembrane subunit ASGP-2 from rat (2 copies).
- Coagulation associated proteins C, Z (2 copies) and S (4 copies).
- Coagulation factors VII, IX, X and XII (2 copies).
- 20 - Complement C1r components (1 copy).
- Complement C1s components (1 copy).
- Complement-activating component of Ra-reactive factor (RARF) (1 copy).
- Complement components C6, C7, C8 alpha and beta chains, and C9 (1 copy).
- Crumbs, an epithelial development protein from Drosophila (29 copies).
- 25 - Epidermal growth factor precursor (7-9 copies).
- Exogastrula-inducing peptides A, C, D and X from sea urchin (1 copy).
- Fat protein, a Drosophila cadherin-related tumor suppressor (5 copies).
- Fetal antigen 1, a probable neuroendocrine differentiation protein, which is derived from the delta-like protein (DLK) (6 copies).
- 30 - Fibrillin 1 (47 copies) and fibrillin 2 (14 copies).
- Fibropellins IA (21 copies), IB (13 copies), IC (8 copies), II (4 copies) and III (8 copies) from the apical lamina - a component of the extracellular matrix - of sea urchin.
- Fibulin-1 and -2, two extracellular matrix proteins (9-11 copies).

- Giant-lens protein (protein Argos), which regulates cell determination and axon guidance in the *Drosophila* eye (1 copy).

- Growth factor-related proteins from various poxviruses (1 copy).

- Gurken protein, a *Drosophila* developmental protein (1 copy).

5 - Heparin-binding EGF-like growth factor (HB-EGF), transforming growth factor alpha (TGF-alpha), growth factors Lin-3 and Spitz (1 copy); the precursors are membrane proteins, the mature form is located extracellular.

- Hepatocyte growth factor (HGF) activator (EC 3.4.21.-) (2 copies).

10 - LDL and VLDL receptors, which bind and transport low-density lipoproteins and very low-density lipoproteins (3 copies).

- LDL receptor-related protein (LRP), which may act as a receptor for endocytosis of extracellular ligands (22 copies).

- Leucocyte antigen CD97 (3 copies), cell surface glycoprotein EMR1 (6 copies) and cell surface glycoprotein F4/80 (7 copies).

15 - Limulus clotting factor C, which is involved in hemostasis and host defense mechanisms in Japanese horseshoe crab (1 copy).

- Meprin A alpha subunit, a mammalian membrane-bound endopeptidase (1 copy).

- Milk fat globule-EGF factor 8 (MFG-E8) from mouse (2 copies).

- Neuregulin GGF-I and GGF-II, two human glial growth factors (1 copy).

20 - Neurexins from mammals (3 copies).

- Neurogenic proteins Notch, Xotch and the human homolog Tan-1 (36 copies), Delta (9 copies) and the similar differentiation proteins Lag-2 from *Caenorhabditis elegans* (2 copies), Serrate (14 copies) and Slit (7 copies) from *Drosophila*.

- Nidogen (also called entactin), a basement membrane protein from chordates (2-6 copies).

25 - Ookinete surface proteins (24 Kd, 25 Kd, 28 Kd) from *Plasmodium* (4 copies).

- Pancreatic secretory granule membrane major glycoprotein GP2 (1 copy).

- Perforin, which lyses non-specifically a variety of target cells (1 copy).

- Proteoglycans aggrecan (1 copy), versican (2 copies), perlecan (at least 2 copies), brevican (1 copy) and chondroitin sulfate proteoglycan (gene PG-M) (2 copies).

30 - Prostaglandin G/H synthase 1 and 2 (EC 1.14.99.1) (1 copy), which is found in the endoplasmic reticulum.

- S1-5, a human extracellular protein whose ultimate activity is probably modulated by the environment (5 copies).

- Schwannoma-derived growth factor (SDGF), an autocrine growth factor as well as a mitogen for different target cells (1 copy).

- Selectins. Cell adhesion proteins such as ELAM-1 (E-selectin), GMP-140 (P-selectin), or the lymph-node homing receptor (L-selectin) (1 copy).

5 - Serine/threonine-protein kinase homolog (gene Pro25) from *Arabidopsis thaliana*, which may be involved in assembly or regulation of light-harvesting chlorophyll A/B protein (2 copies).

- Sperm-egg fusion proteins PH-30 alpha and beta from guinea pig (1 copy).

- Stromal cell derived protein-1 (SCP-1) from mouse (6 copies).

10 - TDGF-1, human teratocarcinoma-derived growth factor 1 (1 copy).

- Tenascin (or neuronectin), an extracellular matrix protein from mammals (14.5 copies), chicken (TEN-A) (13.5 copies) and the related proteins human tenascin-X (18 copies) and tenascin-like proteins TEN-A and TEN-M from *Drosophila* (8 copies).

- Thrombomodulin (fetomodulin), which together with thrombin activates protein C (6 copies).

- Thrombospondin 1, 2 (3 copies), 3 and 4 (4 copies), adhesive glycoproteins that mediate cell-to-cell and cell-to-matrix interactions.

- Thyroid peroxidase 1 and 2 (EC 1.11.1.8) from human (1 copy).

- Transforming growth factor beta-1 binding protein (TGF-B1-BP) (16 or 18 copies).

20 - Tyrosine-protein kinase receptors Tek and Tie (EC 2.7.1.112) (3 copies).

- Urokinase-type plasminogen activator (EC 3.4.21.73) (UPA) and tissue plasminogen activator (EC 3.4.21.68) (TPA) (1 copy).

- Uromodulin (Tamm-horsfall urinary glycoprotein) (THP) (3 copies).

- Vitamin K-dependent anticoagulants protein C (2 copies) and protein S (4 copies) and the similar protein Z, a single-chain plasma glycoprotein of unknown function (2 copies).

- 63 Kd sperm flagellar membrane protein from sea urchin (3 copies).

- 93 Kd protein (gene nel) from chicken (5 copies).

- Hypothetical 337.6 Kd protein T20G5.3 from *Caenorhabditis elegans* (44 copies).

30 The functional significance of EGF domains in what appear to be unrelated proteins is not yet clear. However, a common feature is that these repeats are found in the extracellular domain of membrane-bound proteins or in proteins known to be secreted (exception: prostaglandin G/H synthase). The EGF domain includes six cysteine residues which have been shown (in

730

EGF) to be involved in disulfide bonds. The main structure is a two-stranded beta-sheet followed by a loop to a C-terminal short two-stranded sheet. Subdomains between the conserved cysteines strongly vary in length as shown in the following schematic representation of the EGF-like domain:

```

5      +-----+      +-----+      |      |      |
| x(4)-C-x(0,48)-C-x(3,12)-C-x(1,70)-C-x(1,6)-C-x(2)-G-a-x(0,21)-G-x(2)-C-x  |
|      *****                                         |
|      +-----+                                         |

```

10 'C': conserved cysteine involved in a disulfide bond.

'G': often conserved glycine

'a': often conserved aromatic amino acid

'*': position of both patterns.

'x': any residue

15
20

The region between the 5th and 6th cysteine contains two conserved glycines of which at least one is present in most EGF-like domains. Two patterns were created for this domain, each including one of these C-terminal conserved glycine residues.

Consensus pattern: C-x-C-x(5)-G-x(2)-C [The 3 C's are involved in disulfide bonds]

Sequences known to belong to this class detected by the pattern A majority, but not those that have very long or very short regions between the last 3 conserved cysteines of their EGF-like domain(s). Other sequence(s) detected in SWISS-PROT87 proteins, of which 27 can be considered as possible candidates.

25

Consensus pattern: C-x-C-x(2)-[GP]-[FYW]-x(4,8)-C [The three C's are involved in disulfide bonds] Sequences known to belong to this class detected by the pattern A majority, but not those that have very long or very short regions between the last 3 conserved cysteines of their EGF-like domain(s). Other sequence(s) detected in SWISS-PROT83 proteins, of which 49 can be considered as possible candidates. Note The beta chain of the integrin family of proteins contains 2 cysteine- rich repeats which were said to be dissimilar with the EGF pattern [7].

30

Note Laminin EGF-like repeats (see <PDOC00961>) are longer than the average EGF module and contain a further disulfide bond C-terminal of the EGF-like region. Perlecan and agrin contain both EGF-like domains and laminin-type EGF-like domains. Note the pattern do not detect all of the repeats of proteins with multiple EGF-like repeats. Note see
 5 <PDOC00913> for an entry describing specifically the subset of EGF-like domains that bind calcium.

[1] Davis C.G. New Biol. 2:410-419(1990).

[2] Blomquist M.C., Hunt L.T., Barker W.C. Proc. Natl. Acad. Sci. U.S.A. 81:7363-
 10 7367(1984).

[3] Barker W.C., Johnson G.C., Hunt L.T., George D.G. Protein Nucl. Acid Enz. 29:54-
 68(1986).

[4] Doolittle R.F., Feng D.F., Johnson M.S. Nature 307:558-560(1984).

[5] Appella E., Weber I.T., Blasi F. FEBS Lett. 231:1-4(1988).

[6] Campbell I.D., Bork P. Curr. Opin. Struct. Biol. 3:385-392(1993).

[7] Tamkun J.W., DeSimone D.W., Fonda D., Patel R.S., Buck C., Horwitz A.F., Hynes
 15 R.O. Cell 46:271-282(1986).

20 890. Ham1 family (Ham1p_like)

This family consists of the HAM1 protein Swiss:P47119 and hypothetical archaeal bacterial and C. elegans proteins. HAM1 controls 6-N-hydroxylaminopurine (HAP) sensitivity and mutagenesis in S. cerevisiae Swiss:P47119 [1]. The HAM1 protein protects the cell from HAP, either on the level of deoxynucleoside triphosphate or the DNA level by a yet
 25 unidentified set of reactions [1]. Number of members: 19

[1] Noskov VN, Staak K, Shcherbakova PV, Kozmin SG, Negishi K, Ono BC, Hayatsu H, Pavlov YI; Medline: 96381244 "HAM1, the gene controlling 6-N-hydroxylaminopurine sensitivity and mutagenesis in the yeast Saccharomyces cerevisiae." Yeast 1996;12:17-29.

30 891. (HCO3_cotransp)

Anion exchange is a cellular transport function which contributes to the regulation of cell pH and volume. Anion exchangers are a family of functionally related proteins that contributes to these properties by maintaining the intracellular level of the two principal anions: chloride and HCO₃⁻. The best characterized anion exchanger is the band 3 protein [1], which is an erythrocyte anion exchange membrane glycoprotein. Band 3 is a protein of about 900 amino acids which consists of a cytoplasmic N-terminal domain of about 400 residues and an hydrophobic C-terminal section of about 500 residues that contains at least ten transmembrane regions. The cytoplasmic domain provides binding sites for cytoskeletal proteins, while the integral membrane domain is responsible for anion transport. Band 3 protein is specific to erythroid cells, at least two other proteins [2] structurally and functionally related to band 3, are found in nonerythroid tissues:

- AE2 (or B3 related protein; B3RP), a protein of 1200 residues, which seems to be present in a variety of cell types including lymphoid, kidney, and choroid plexus.
- AE3, a protein of 1200 residues, which is specific to neurons.

Structurally AE2 and AE3 are very similar to band 3, the main difference being an extension of some 300 residues of the N-terminal domain in AE2 and AE3.

Two signature patterns were developed for these proteins. The first pattern is based on a conserved stretch of sequence that contains four clustered positive charged residues and which is located at the C-terminal extremity of the cytoplasmic domain, just before the first transmembrane segment from the integral domain. The second pattern is based on the perfectly conserved sequence of the fifth transmembrane segment; this segment contains a lysine, which is the covalent binding site for the isothiocyanate group of DIDS, an inhibitor of anion exchange.

Consensus pattern F-G-G-[LIVM](2)-[KR]-D-[LIVM]-[RK]-R-R-Y Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern [FI]-L-I-S-L-I-F-I-Y-E-T-F-x-K-L Sequences known to belong to this class detected by the pattern ALL.

[1] Jay D., Cantley L. Annu. Rev. Biochem. 55:511-538(1986).

[2] Reithmeier R.A.F. Curr. Opin. Struct. Biol. 3:515-523(1993).

892. ATP phosphoribosyltransferase signature (HisG)

ATP phosphoribosyltransferase (EC 2.4.2.17) is the enzyme that catalyzes the first step in the biosynthesis of histidine in bacteria, fungi and plants. It is a protein of about 23 to 32 Kd. As a signature pattern a region located in the C-terminal part of this enzyme was selected.

Consensus pattern E-x(5)-G-x-[SAG]-x(2)-[IV]-x-D-[LIV]-x(2)-[ST]-G-x-T-[LM]

Sequences known to belong to this class detected by the pattern ALL.

10

893. HNH endonuclease (HNH)

Number of members: 56

[1] Shub DA, Goodrich-Blair H, Eddy SR; Medline: 95117127 "Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns."

Trends Biochem Sci 1994;19:402-404.

[2] Dalgaard JZ, Klar AJ, Moser MJ, Holley WR, Chatterjee A, Mian IS; Medline: 98026854 "Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family." Nucleic Acids Res 1997;25:4626-4638.

[3] Gorbalenya AE; Medline: 95004046 "Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family." Protein Sci 1994;3:1117-1120.

894. NEUROHYPOPHYS_HORM (hormone5)

Oxytocin (or ocytocin) and vasopressin [1] are small (nine amino acid residues), structurally and functionally related neurohypophysial peptide hormones. Oxytocin causes contraction of the smooth muscle of the uterus and of the mammary gland while vasopressin has a direct antidiuretic action on the kidney and also causes vasoconstriction of the peripheral vessels.

Like the majority of active peptides, both hormones are synthesized as larger protein precursors that are enzymatically converted to their mature forms. Peptides belonging to this family are also found in birds, fish, reptiles and amphibians (mesotocin, isotocin, valitocin, glutitocin, aspartocin, vasotocin, seritocin, asvatocin, phasvatocin), in worms (anetocin), octopi (cephalotocin), locust (locupressin or neuropeptide F1/F2) and in molluscs

(conopressins G and S) [2]. The pattern developed to detect this category of peptides spans their entire sequence and includes four invariant amino acid residues.

Consensus pattern C-[LIFY](2)-x-N-[CS]-P-x-G [The two C's are linked by a disulfide bond]. Sequences known to belong to this class detected by the pattern ALL.

[1] Acher R., Chauvet J. Biochimie 70:1197-1207(1988).

[2] Chauvet J., Michel G., Ouedraogo Y., Chou J., Chait B.T., Acher R. Int. J. Pept. Protein Res. 45:482-487(1995).

895. 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (HPPK)

All organisms require reduced folate cofactors for the synthesis of a variety of metabolites. Most microorganisms must synthesize folate de novo because they lack the active transport system of higher vertebrate cells which allows these organisms to use dietary folates.

Enzymes involved in folate biosynthesis are therefore targets for a variety of antimicrobial agents such as trimethoprim or sulfonamides. 7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (EC 2.7.6.3) (HPPK) catalyzes the attachment of pyrophosphate to 6-hydroxymethyl-7,8-dihydropterin to form 6-hydroxymethyl-7,8-dihydropteridine pyrophosphate. This is the first step in a three-step pathway leading to 7,8-dihydrofolate.

Bacterial HPPK (gene folK or sulD) [1] is a protein of 160 to 270 amino acids. In the lower eukaryote *Pneumocystis carinii*, HPPK is the central domain of a multifunctional folate synthesis enzyme (gene fas) [2]. As a signature for HPPK, a conserved region located in the central section of these enzymes was selected.

Consensus pattern [KRHD]-x-[GA]-[PSAE]-R-x(2)-D-[LIV]-D-[LIVM](2) Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT/NONE.

[1] Talarico T.L., Ray P.H., Dev I.K., Merrill B.M., Dallas W.S. J. Bacteriol. 174:5971-5977(1992).

[2] Volpes F., Dyer M., Scaife J.G., Darby G., Stammers D.K., Delves C.J. Gene 112:213-218(1992).

896. Metalloenzyme superfamily (Metalloenzyme)

This family includes phosphopentomutase Swiss:P07651 and 2,3-bisphosphoglycerate-independent phosphoglycerate mutase, Swiss:P37689. This family is also related to alk_phosphatase [1]. The alignment contains the most conserved residues that are probably involved in metal binding and catalysis. Number of members: 34

[1] Galperin MY, Bairoch A, Koonin EV; Medline: 99180418 "A superfamily of metalloenzymes unifies phosphopentomutase and cofactor-independent phosphoglycerate mutase with alkaline phosphatases and sulfatases." Protein Sci 1998;7:1829-1835.

897. Penicillin amidase (Penicil_amidase)

Penicillin amidase or penicillin acylase EC:3.5.1.11 catalyses the hydrolysis of benzylpenicillin to phenylacetic acid and 6-aminopenicillanic acid (6-APA) a key intermediate in the the synthesis of penicillins [1]. Also in the family is cephalosporin acylase Swiss:P07662 and Swiss:P29958 aculeacin A acylase which are involved in the synthesis of related peptide antibiotics. Number of members: 13

[1] Verhaert RM, Riemens AM, van der Laan JM, van Duin J, Quax WJ; Medline: 97438505 "Molecular cloning and analysis of the gene encoding the thermostable penicillin G acylase from *Alcaligenes faecalis*. Appl Environ Microbiol 1997;63:3412-3418.

[2] Duggleby HJ, Tolley SP, Hill CP, Dodson EJ, Dodson G, Moody PC; Medline: 95115804 "Penicillin acylase has a single-amino-acid catalytic centre." Nature 1995;373:264-268.

898. Phosphoribosyl-AMP cyclohydrolase (PRA-CH)

This enzyme catalyses the third step in the histidine biosynthetic pathway. It requires Zn ions for activity. Number of members: 13

[1] D'Ordine RL, Klem TJ, Davisson VJ; Medline: 99129952 "N1-(5'-phosphoribosyl)adenosine-5'-monophosphate cyclohydrolase: purification and characterization of a unique metalloenzyme. *Biochemistry* 1999;38:1537-1546.

5

899. Phosphoribosyl-ATP pyrophosphohydrolase (PRA-PH)

This enzyme catalyses the second step in the histidine biosynthetic pathway. Number of members: 32

10

[1] Keesey JK Jr, Bigelis R, Fink GR; Medline: 79216449 "The product of the *his4* gene cluster in *Saccharomyces cerevisiae*. A trifunctional polypeptide." *J Biol Chem* 1979 Aug 10;254:7427-7433.

[2] Bruni CB, Carlomagno MS, Formisano S, Paoletta G; Medline: 86310274 "Primary and secondary structural homologies between the *HIS4* gene product of *Saccharomyces cerevisiae* and the *hisIE* and *hisD* gene products of *Escherichia coli* and *Salmonella typhimurium*." *Mol Gen Genet* 1986;203:389-396.

5

900. Prokaryotic membrane lipoprotein lipid attachment site (PstS)

In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):

20

- Major outer membrane lipoprotein (murein-lipoproteins) (gene *lpp*).
- *Escherichia coli* lipoprotein-28 (gene *nlpA*).
- *Escherichia coli* lipoprotein-34 (gene *nlpB*).
- *Escherichia coli* lipoprotein *nlpC*.
- *Escherichia coli* lipoprotein *nlpD*.

25

- *Escherichia coli* osmotically inducible lipoprotein B (gene *osmB*).
- *Escherichia coli* osmotically inducible lipoprotein E (gene *osmE*).
- *Escherichia coli* peptidoglycan-associated lipoprotein (gene *pal*).
- *Escherichia coli* rare lipoproteins A and B (genes *rplA* and *rplB*).

30

- Escherichia coli copper homeostasis protein cutF (or nlpE).
- Escherichia coli plasmids traT proteins.
- Escherichia coli Col plasmids lysis proteins.
- A number of Bacillus beta-lactamases.
- 5 - Bacillus subtilis periplasmic oligopeptide-binding protein (gene oppA).
- Borrelia burgdorferi outer surface proteins A and B (genes ospA and ospB).
- Borrelia hermsii variable major protein 21 (gene vmp21) and 7 (gene vmp7).
- Chlamydia trachomatis outer membrane protein 3 (gene omp3).
- Fibrobacter succinogenes endoglucanase cel-3.
- 10 - Haemophilus influenzae proteins Pal and Pcp.
- Klebsiella pullulunase (gene pulA).
- Klebsiella pullulunase secretion protein pulS.
- Mycoplasma hyorhinitis protein p37.
- Mycoplasma hyorhinitis variant surface antigens A, B, and C (genes vlpABC).
- 15 - Neisseria outer membrane protein H.8.
- Pseudomonas aeruginosa lipopeptide (gene lppL).
- Pseudomonas solanacearum endoglucanase egl.
- Rhodopseudomonas viridis reaction center cytochrome subunit (gene cytC).
- Rickettsia 17 Kd antigen.
- 20 - Shigella flexneri invasion plasmid proteins mxiJ and mxiM.
- Streptococcus pneumoniae oligopeptide transport protein A (gene amiA).
- Treponema pallidum 34 Kd antigen.
- Treponema pallidum membrane protein A (gene tmpA).
- Vibrio harveyi chitinase (gene chb).

- 25 - Yersinia virulence plasmid protein yscJ.
 - Halocyanin from Natrobacterium pharaonis [4], a membrane associated copper-binding protein. This is the first archaebacterial protein known to be modified in such a fashion).
- From the precursor sequences of all these proteins, a consensus pattern was derived and a set of rules to identify this type of post-translational modification.

30 Consensus pattern {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1) The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least one Lys or one Arg in the first

seven positions of the sequence. Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROT some 100 prokaryotic proteins. Some of them are not membrane lipoproteins, but at least half of them could be.

- 5 [1] Hayashi S., Wu H.C. J. Bioenerg. Biomembr. 22:451-471(1990).
 [2] Klein P., Somorjai R.L., Lau P.C.K. Protein Eng. 2:15-20(1988).
 [3] von Heijne G. Protein Eng. 2:531-534(1989).
 [4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).

10

901. Ribosome recycling factor (RRF)

The ribosome recycling factor (RRF / ribosome release factor) dissociates the ribosome from the mRNA after termination of translation, and is essential bacterial growth [1]. Thus ribosomes are "recycled" and ready for another round of protein synthesis. Number of members: 27

- [1] Janosi L, Shimizu I, Kaji A; Medline: 94240115 "Ribosome recycling factor (ribosome releasing factor) is essential for bacterial growth." Proc Natl Acad Sci U S A 1994;91:4249-4253.

902. S-layer homology(SLH)

S-layers are paracrystalline mono-layered assemblies of (glyco)proteins which coat the surface of bacteria [1]. Several S-layer proteins and some other cell wall proteins contain one or more copies of a domain of about 50-60 residues, which has been called SLH (for S-layer homology) [2]. There is strong evidence that this domain serves as an anchor to the peptidoglycan [3]. The SLH domain has been found in:

- S-layer glycoprotein of *Acetogenium kivui* (3 copies).
- S-layer 125 Kd protein of *Bacillus sphaericus* (3 copies).
- S-layer protein of *Bacillus anthracis* (3 copies).
- S-layer protein of *Bacillus licheniformis* (3 copies).
- S-layer protein (HWP) from *Bacillus brevis* strain HPD31 (3 copies).

- Middle cell wall protein (MWP) from *Bacillus brevis* strain 47 (3 copies).
- S-layer protein (p100) of *Thermus thermophilus* (1 copy).
- Outer membrane protein Omp-alpha from *Thermotoga maritima* (1 copy).
- Cellulosome anchoring protein (gene *ancA*), outer layer protein B (OlpB) and a further
5 potential cell surface glycoprotein from *Clostridium thermocellum* (3 copies; the first copy is
missing its N-terminal third which is appended to the end of the third copy; may have arisen
by circular permutation).
- Amylopullulanase (gene *amyB*) from *Thermoanaerobacter thermosulfurogenes* (3 copies)
- Amylopullulanase (gene *aapT*) from *Bacillus* strain XAL-601 (3 copies).
- 10 - Endoglucanase from *Bacillus* strain KSM-635 (3 copies).
- Exoglucanase (gene *xynX*) from *Clostridium thermocellum* (3 copies).
- Xylanase A (gene *xynA*) from *Thermoanaerobacter saccharolyticum* (2 copies; 3 copies if a
frameshift is taken into account).
- Protein involved in butirosin production (ButB) from *Bacillus circulans* (2 incomplete
15 copies; 3 copies if three frameshifts are taken into account).
- Two hypothetical proteins from *Synechocystis* strain PCC 6803 (1 copy each).
- A hypothetical protein with sequence similarity to amylopullulanases found 3' of amylase
gene from *Bacillus circulans* (fragment of 1 copy; 3 copies if two frameshifts are taken into
account).
- 20 SLH domains are found at the N- or C-termini of mature proteins. They occur in single copy
followed by a predicted coiled coil domain, or in three contiguous copies. Structurally, the
SLH domain is predicted to contain two alpha-helices flanking a beta strand. The SLH
sequences are fairly divergent with an average identity of about 25%. It is however possible
to build a sequence pattern that starts at the second position of the domain and that spans 3/4
25 of its length.

Consensus pattern[LVFYT]-x-[DA]-x(2,5)-[DNGSATPHY]-[FYWPDA]-x(4)-[LIV]-x(2)-
[GTALV]-x(4,6)-[LIVFYC]-x(2)-G-x-[PGSTA]-x(2,3)-[MFYA]-x- [PGAV]-x(3,10)-
[LIVMA]-[STKR]-[RY]-x-[EQ]-x-[STALIVM] Sequences known to belong to this class
30 detected by the pattern ALL. Other sequence(s) detected in SWISS-PROTNONE.

[1] Beveridge T.J. Curr. Opin. Struct. Biol. 4:204-212(1994).

[2] Lupas A., Engelhardt H., Peters J., Santarius U., Volker S., Baumeister W. *J. Bacteriol.* 176:1224-1233(1994).

[3] Lemaire M., Ohayon H., Gounon P., Fujino T., Beguin P. *J. Bacteriol.* 177:2451-2459(1995).

5

903. Queuine tRNA-ribosyltransferase (TGT)

This is a family of queuine tRNA-ribosyltransferases EC:2.4.2.29, also known as tRNA-guanine transglycosylase and guanine insertion enzyme. Queuine tRNA-ribosyltransferase modifies tRNAs for asparagine, aspartic acid, histidine and tyrosine with queuine. It catalyses the exchange of guanine-34 at the wobble position with 7-aminomethyl-7-deazaguanine, and the addition of a cyclopentenediol moiety to 7-aminomethyl-7-deazaguanine-34 tRNA; giving a hypermodified base queuine in the wobble position [1,2]. The aligned region contains a zinc binding motif C-x-C-x2-C-x29-H, and important tRNA and 7-aminomethyl-7-deazaguanine binding residues [1]. Number of members: 27

[1] Romier C, Reuter K, Suck D, Ficner R; Medline: 96256303 "Crystal structure of tRNA-guanine transglycosylase: RNA modification by base exchange." *EMBO J* 1996;15:2850-2857.

[2] Garcia GA, Koch KA, Chong S; Medline: 93287116 "tRNA-guanine transglycosylase from *Escherichia coli*. Overexpression, purification and quaternary structure." *J Mol Biol* 1993;231:489-497.

904. ThiC Family (ThiC)

ThiC is found within the thiamine biosynthesis operon. ThiC is involved in pyrimidine biosynthesis [2]. ThiC catalyzes the substitution of the pyrophosphate of 2-methyl-4-amino-5-hydroxymethylpyrimidine pyrophosphate by 4-methyl-5-(beta-hydroxyethyl)thiazole phosphate to yield thiamine phosphate [3]. Number of members: 12

[1] Vander Horn PB, Backstrom AD, Stewart V, Begley TP; Medline: 93163063 "Structural genes for thiamine biosynthetic enzymes (thiCEFGH) in *Escherichia coli* K-12." *J Bacteriol* 1993;175:982-992.

[2] Begley TP, Downs DM, Ealick SE, McLafferty FW, Van Loon AP, Taylor S, Campobasso N, Chiu HJ, Kinsland C, Reddick JJ, Xi J; Medline: 99311269 "Thiamin biosynthesis in prokaryotes." *Arch Microbiol* 1999;171:293-300.

[3] Zhang Y, Taylor SV, Chiu HJ, Begley TP; Medline: 97284509 "Characterization of the *Bacillus subtilis* thiC operon involved in thiamine biosynthesis." *J Bacteriol* 1997;179:3030-3035.

905. Putative tRNA binding domain (tRNA_bind)

This domain is found in prokaryotic methionyl-tRNA synthetases, prokaryotic phenylalanyl tRNA synthetases the yeast GU4 nucleic-binding protein (G4p1 or p42, ARC1) [2], human tyrosyl-tRNA synthetase [1], and endothelial-monocyte activating polypeptide II. G4p1 binds specifically to tRNA form a complex with methionyl-tRNA synthetases [2]. In human tyrosyl-tRNA synthetase this domain may direct tRNA to the active site of the enzyme [2].

This domain may perform a

common function in tRNA aminoacylation [1]. Number of members: 12

[1] Kleeman TA, Wei D, Simpson KL, First EA; Medline: 97306356 "Human tyrosyl-tRNA synthetase shares amino acid sequence homology with a putative cytokine." *J Biol Chem* 1997;272:14420-14425.

[2] Simos G, Segref A, Fasiolo F, Hellmuth K, Shevchenko A, Mann M, Hurt EC; Medline: 97050848 "The yeast protein Arc1p binds to tRNA and functions as a cofactor for the methionyl-and glutamyl-tRNA synthetases." *EMBO J* 1996;15:5437-5448.

906. UbiA prenyltransferase family signature (UbiA)

The following prenyltransferases are evolutionary related [1,2]:

- Bacterial 4-hydroxybenzoate octaprenyltransferase (gene ubiA).
- Yeast mitochondrial para-hydroxybenzoate--polyprenyltransferase (gene COQ2).
- Protoheme IX farnesyltransferase (heme O synthase) from yeast and mammals (gene COX10) and from bacteria (genes cyoE or ctaB).

These proteins probably contain seven transmembrane segments. The best conserved region is located in a loop between the second and third of these segments and was used as a signature pattern.

5 Consensus pattern N-x(3)-[DE]-x(2)-[LIF]-D-x(2)-[VM]-x-R-[ST]-x(2)-R-x(4)-G Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.

[1] Melzer M., Heide L. Biochim. Biophys. Acta 1212:93-102(1994).

10 [2] Mogi T., Saiki K., Anraku Y. Mol. Microbiol. 14:391-398(1994).

907. Uncharacterized protein family UPF0044 signature (UPF0044)

The following uncharacterized proteins have been shown [1] to be highly similar:

- Bacillus subtilis hypothetical protein yqeI.
- Escherichia coli hypothetical protein yhbY and HI1333, the corresponding Haemophilus influenzae protein.
- Methanococcus jannaschii hypothetical protein MJ0652.

These are small proteins of 10 to 15 Kd. They can be picked up in the database by the following pattern. This pattern is located in the N-terminal part of these proteins.

Consensus pattern L-[ST]-x(3)-K-x(3)-[KR]-[SGA]-x-[GA]-H-x-L-x-P-[LIV]-x(2)-[LIV]-[GA]-x(2)-G Sequences known to belong to this class detected by the pattern ALL.

25

908. ATP synthase (C/AC39) subunit (vATP-synt_AC39)

This family includes the AC39 subunit from vacuolar ATP synthase Swiss:P32366 [1], and the C subunit from archaeobacterial ATP synthase [2]. The family also includes subunit C from the Sodium transporting ATP synthase from Enterococcus hirae Swiss:P43456 [3].

30 Number of members: 12

10

15

20

25

30

The WW domain [1-4,E1] (also known as *rsp5* or WWP) has been originally discovered as a short conserved region in a number of unrelated proteins, among them dystrophin, the gene responsible for Duchenne muscular dystrophy. The domain, which spans about 35 residues, is repeated up to 4 times in some proteins. It has been shown [5] to bind proteins with particular proline- motifs, [AP]-P-P-[AP]-Y, and thus resembles somewhat SH3 domains. It appears to contain beta-strands grouped around four conserved aromatic positions; generally Trp. The name WW or WWP derives from the presence of these Trp as well as that of a conserved Pro. It is frequently associated with other domains typical for proteins in signal transduction processes.

Proteins containing the WW domain are listed below.

- Dystrophin, a multidomain cytoskeletal protein. Its longest alternatively spliced form consists of an N-terminal actin-binding domain, followed by 24 spectrin-like repeats, a cysteine-rich calcium-binding domain and a C-terminal globular domain. Dystrophin forms tetramers and is thought to have multiple functions including involvement in membrane stability, transduction of contractile forces to the extracellular environment and organization of membrane specialization. Mutations in the dystrophin gene lead to muscular dystrophy of Duchenne or Becker type. Dystrophin contains one WW domain C-terminal of the spectrin-repeats.
- Utrophin, a dystrophin-like protein of unknown function.
- Vertebrate YAP protein is a substrate of an unknown serine kinase. It binds to the SH3 domain of the Yes oncoprotein via a proline-rich region. This protein appears in alternatively spliced isoforms, containing either one or two WW domains [6].
- Mouse NEDD-4 plays a role in the embryonic development and differentiation of the central nervous system. It contains 3 WW modules followed by a HECT domain. The human ortholog contains 4 WW domains, but the third WW domain is probably spliced resulting in an alternate NEDD-4 protein with only 3 WW modules [3].
- Yeast RSP5 is similar to NEDD-4 in its molecular organization. It contains an N-terminal C2 domain (see <PDOC00380>, followed by a histidine-rich region, 3 WW domains and a HECT domain.
- Rat FE65, a transcription-factor activator expressed preferentially in liver. The activator domain is located within the N-terminal 232 residues of FE65, which also contain the WW domain.
- Yeast ESS1/PTF1, a putative peptidyl prolyl cis-trans isomerase from family ppiC (see <PDOC00840>). A related protein, dodo (gene dod) exists in Drosophila and in mammals (gene PIN1).
- Tobacco DB10 protein. The WW domain is located N-terminal to the region with similarity to ATP-dependent RNA helicases.
- IQGAP, a human GTPase activating protein acting on ras. It contains an N-terminal domain similar to fly muscle mp20 protein and a C-terminal ras GTPase activator domain.
- Yeast pre-mRNA processing protein PRP40, Caenorhabditis elegans ZK1098.1 and fission yeast SpAC13C5.02 are related proteins with similarity to MYO2-type myosin, each containing two WW-domains at the N-terminus.

745

- *Caenorhabditis elegans* hypothetical protein C38D4.5, which contains one WW module, a PH domain (see <PDOC50003>) and a C-terminal phosphatidylinositol 3-kinase domain.

- Yeast hypothetical protein YFL010c.

For the sensitive detection of WW domains, a profile was developed which spans the whole
5 homology region as well as a pattern.

Consensus pattern W-x(9,11)-[VFY]-[FYW]-x(6,7)-[GSTNE]-[GSTQCR]-[FYW]-x(2)-P
Sequences known to belong to this class detected by the pattern ALL. Other sequence(s)
detected in SWISS-PROT8. Sequences known to belong to this class detected by the
10 profileALL.

[1] Bork P., Sudol M. Trends Biochem. Sci. 19:531-533(1994).

[2] Andre B., Springael J.Y. Biochem. Biophys. Res. Commun. 205:1201-1205(1994).

[3] Hofmann K.O., Bucher P. FEBS Lett. 358:153-157(1995).

[4] Sudol M., Chen H.I., Bougeret C., Einbond A., Bork P. FEBS Lett. 369:67-71(1995).

[5] Chen H.I., Sudol M. Proc. Natl. Acad. Sci. U.S.A. 92:7819-7823(1995).

[6] Sudol M., Bork P., Einbond A., Kastury K., Druck T., Negrini M., Huebner K., Lehman
D. J. Biol. Chem. 270:14733-14741(1995).

911. Xeroderma pigmentosum (XP) [1] (XPG_1)

Xeroderma pigmentosum (XP) [1] is a human autosomal recessive disease, characterized by a
high incidence of sunlight-induced skin cancer. People's skin cells with this condition are
hypersensitive to ultraviolet light, due to defects in the incision step of DNA excision repair.

25 There are a minimum of seven genetic complementation groups involved in this pathway:
XP-A to XP-G. The defect in XP-G can be corrected by a 133 Kd nuclear protein called XPG
(or XPGC) [2].

XPG belongs to a family of proteins [2,3,4,5,6] that are composed of two main subsets:

30 - Subset 1, to which belongs XPG, RAD2 from budding yeast and rad13 from fission yeast.
RAD2 and XPG are single-stranded DNA endonucleases [7,8]. XPG makes the 3'incision in
human DNA nucleotide excision repair [9].

- Subset 2, to which belongs mouse and human FEN-1, rad2 from fission yeast, and RAD27 from budding yeast. FEN-1 is a structure-specific endonuclease.

In addition to the proteins listed in the above groups, this family also includes:

- 5 - Fission yeast exo1, a 5'->3' double-stranded DNA exonuclease that could act in a pathway that corrects mismatched base pairs.
- Yeast EXO1 (DHS1), a protein with probably the same function as exo1.
- Yeast DIN7.

10 Sequence alignment of this family of proteins reveals that similarities are largely confined to two regions. The first is located at the N-terminal extremity (N-region) and corresponds to the first 95 to 105 amino acids. The second region is internal (I-region) and found towards the C-terminus; it spans about 140 residues and contains a highly conserved core of 27 amino acids that includes a conserved pentapeptide (E-A-[DE]-A-[QS]). It is possible that the conserved acidic residues are involved in the catalytic mechanism of DNA excision repair in XPG. The amino acids linking the N- and I-regions are not conserved; indeed, they are largely absent from proteins belonging to the second subset.

20 Two signature patterns were developed for these proteins. The first corresponds to the central part of the N-region, the second to part of the I-region and includes the putative catalytic core pentapeptide.

25 Consensus pattern [VI]-[KRE]-P-x-[FYIL]-V-F-D-G-x(2)-[PIL]-x-[LVC]-K Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROTNONE.

30 Consensus pattern [GS]-[LIVM]-[PER]-[FYS]-[LIVM]-x-A-P-x-E-A-[DE]-[PAS]- [QS]-[CLM] Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROTNONE.

[1] Tanaka K., Wood R.D. Trends Biochem. Sci. 19:83-86(1994).

[2] Scherly D., Nospikel T., Corlet J., Ucla C., Bairoch A., Clarkson S.G. Nature 363:182-185(1993).

- [3] Carr A.M., Sheldrick K.S., Murray J.M., Al-Harithy R., Watts F.Z., Lehmann A.R. Nucleic Acids Res. 21:1345-1349(1993).
- [4] Murray J.M., Tavassoli M., Al-Harithy R., Sheldrick K.S., Lehmann A.R., Carr A.M., Watts F.Z. Mol. Cell. Biol. 14:4878-4888(1994).
- 5 [5] Harrington J.J., Lieber M.R. Genes Dev. 8:1344-1355(1994).
- [6] Szankasi P., Smith G.R. Science 267:1166-1169(1995).
- [7] Habraken Y., Sung P., Prakash L., Prakash S. Nature 366:365-368(1993).
- [8] O'Donovan A., Scherly D., Clarkson S.G., Wood R.D. J. Biol. Chem. 269:15965-15968(1994).
- 10 [9] O'Donovan A., Davies A.A., Moggs J.G., West S.C., Wood R.D. Nature 371:432-435(1994).

912. 5-formyltetrahydrofolate cyclo-ligase (5-FTHF_cyc-lig)

5-formyltetrahydrofolate cyclo-ligase or methenyl-THF synthetase EC:6.3.3.2 catalyses the interchange of 5-formyltetrahydrofolate (5-FTHF) to 5-10-methenyltetrahydrofolate, this requires ATP and Mg²⁺ [1]. 5-FTHF is used in chemotherapy where it is clinically known as Leucovorin [2].

Number of members: 23

- [1] Dayan A, Bertrand R, Beauchemin M, Chahla D, Mamo A, Filion M, Skup D, Massie B, Jolivet J; Medline: 96096540 "Cloning and characterization of the human 5,10-methenyltetrahydrofolate synthetase-encoding cDNA." Gene 1995;165:307-311.
- 25 [2] Maras B, Stover P, Valiante S, Barra D, Schirch V; Medline: 94308074 "Primary structure and tetrahydropteroylglutamate binding site of rabbit liver cytosolic 5,10-methenyltetrahydrofolate synthetase." J Biol Chem 1994;269:18429-18433.

913. Cytosolic long-chain acyl-CoA thioester hydrolase (Acyl-CoA_hydro)

This family consist of various cytosolic long-chain acyl-CoA thioester hydrolases including human and rat [1,2]. The aligned region is repeated with in the sequence of human and rat cytosolic long-chain acyl-CoA thioester hydrolases of this family. Long-chain acyl-CoA

hydrolases hydrolyse palmitoyl-CoA to CoA and palmitate, they also catalyse the hydrolysis of other long chain fatty acyl-CoA thioesters. Long-chain acyl-CoA hydrolases are present in all living organisms and they may provide a mechanism for the control of lipid metabolism [1].

5 Number of members: 24

[1]Yamada J, Furihata T, Iida N, Watanabe T, Hosokawa M, Satoh T, Someya A, Nagaoka I, Suga T; Medline: 97236308 "Molecular cloning and expression of cDNAs encoding rat brain and liver cytosolic long-chain acyl-CoA hydrolases." Biochem Biophys Res Commun 10 1997;232:198-203.

[2] Broustas CG, Larkins LK, Uhler MD, Hajra AK; Medline: 96209964 "Molecular cloning and expression of cDNA encoding rat brain cytosolic acyl-coenzyme A thioester hydrolase." J Biol Chem 1996;271:10470-10476.

15 914. Agglutinin

Lectin (probable mannose binding)

Members of this family are plant lectins. Many if not all are mannose specific.

Number of members: 87

20 [1] Wright CS, Hester G; Medline: 97094989 "The 2.0 Å structure of a cross-linked complex between snowdrop lectin and a branched mannopentaose: evidence for two unique binding modes." Structure 1996;4:1339-1352.

25 915. (ANF_RECEPTORS)

Natriuretic peptides are hormones involved in the regulation of fluid and electrolyte homeostasis. These hormones stimulate the intracellular production of cyclic GMP as a second messenger.

30 Currently, three types of natriuretic peptide receptors are known [1,2]. Two express guanylate cyclase activity: GC-A (or ANP-A) which seems specific to atrial natriuretic peptide (ANP), and GC-B (or ANP-B) which seems to be stimulated more effectively by brain natriuretic

peptide (BNP) than by ANP. The third receptor (ANP-C) is probably responsible for the clearance of ANP from the circulation and does not play a role in signal transduction.

GC-A and GC-B are plasma membrane-bound proteins that share the following topology: an N-terminal extracellular domain which acts as the ligand binding region, then a transmembrane domain followed by a large cytoplasmic C-terminal region that can be subdivided into two domains: a protein kinase-like domain (see <PDOC00100>) that appears important for proper signalling and a guanylate cyclase catalytic domain (see <PDOC00425>). The topology of ANP-C is different: like GC-A and -B it possesses an extracellular ligand-binding region and a transmembrane domain, but its cytoplasmic domain is very short.

A pattern was developed from the ligand-binding region of natriuretic peptide receptors based on a highly conserved region located in the N-terminal part of the domain.

Consensus pattern G-P-x-C-x-Y-x-A-A-x-V-x-R-x(3)-H-W Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.

[1] Garbers D.L. New Biol. 2:499-504(1990).

[2] Schulz S., Chinkers M., Garbers D.L. FASEB J. 2:2026-2035(1989).

916. (Apocytochrome)

Cytochrome c family heme-binding site signature

In proteins belonging to cytochrome c family [1], the heme group is covalently attached by thioether bonds to two conserved cysteine residues. The consensus sequence for this site is Cys-X-X-Cys-His and the histidine residue is one of the two axial ligands of the heme iron. This arrangement is shared by all proteins known to belong to cytochrome c family, which presently includes cytochromes c, c', c1 to c6, c550 to c556, cc3/Hmc, cytochrome f and reaction center cytochrome c.

750

Consensus pattern C-{CPWHF}-{CPWR}-C-H-{CFYW} Sequences known to belong to this class detected by the pattern ALL, except for four cytochrome c's which lack the first thioether bond. Other sequence(s) detected in SWISS-PROT 454.

- 5 Note: some cytochrome c's have more than a single bound heme group c4 has 2, c7 has 3, c3 has 4, the reaction center has 4, and cc3/Hmc has 16 !

[1] Mathews F.S. Prog. Biophys. Mol. Biol. 45:1-56(1985).

- 10 917. ATP-synt_A-c. ATP synthase Alpha chain, C terminal

[1] Medline: 94344236. Structure at 2.8 A resolution of F1-ATPase from bovine heart mitochondria. Abrahams JP, Leslie AG, Lutter R, Walker JE; Nature 1994;370:621-628.

Number of members: 125

- 15 918. (Basic)

Myc-type, 'helix-loop-helix' dimerization domain signature

HELIX_LOOP_HELIX

- 20 A number of eukaryotic proteins, which probably are sequence specific DNA- binding proteins that act as transcription factors, share a conserved domain of 40 to 50 amino acid residues. It has been proposed [1] that this domain is formed of two amphipathic helices joined by a variable length linker region that could form a loop. This 'helix-loop-helix' (HLH) domain mediates protein dimerization and has been found in the proteins listed below [2,3,E1,E2]. Most of these proteins have an extra basic region of about 15 amino acid
- 25 residues that is adjacent to the HLH domain and specifically binds to DNA. They are referred as basic helix-loop-helix proteins (bHLH), and are classified in two groups: class A (ubiquitous) and class B (tissue-specific). Members of the bHLH family bind variations on the core sequence 'CANNTG', also referred to as the E-box motif. The homo- or heterodimerization mediated by the HLH domain is independent of, but necessary for DNA
- 30 binding, as two basic regions are required for DNA binding activity. The HLH proteins lacking the basic domain (Emc, Id) function as negative regulators since they form heterodimers, but fail to bind DNA. The hairy-related proteins (hairy, E(spl), deadpan) also

repress transcription although they can bind DNA. The proteins of this subfamily act together with co-repressor proteins, like groucho, through their C-terminal motif WRPW.

- The myc family of cellular oncogenes [4], which is currently known to contain four members: c-myc [E3], N-myc, L-myc, and B-myc. The myc genes are thought to play a role in cellular differentiation and proliferation.

- Proteins involved in myogenesis (the induction of muscle cells). In mammals MyoD1 (Myf-3), myogenin (Myf-4), Myf-5, and Myf-6 (Mrf4 or herculin), in birds CMD1 (QMF-1), in *Xenopus* MyoD and MF25, in *Caenorhabditis elegans* CeMyoD, and in *Drosophila* nautilus (nau).

- Vertebrate proteins that bind specific DNA sequences ('E boxes') in various immunoglobulin chains enhancers: E2A or ITF-1 (E12/pan-2 and E47/pan-1), ITF-2 (tcf4), TFE3, and TFEB.

- Vertebrate neurogenic differentiation factor 1 that acts as differentiation factor during neurogenesis.

- Vertebrate MAX protein, a transcription regulator that forms a sequence- specific DNA-binding protein complex with myc or mad.

- Vertebrate Max Interacting Protein 1 (MXI1 protein) which acts as a transcriptional repressor and may antagonize myc transcriptional activity by competing for max.

- Proteins of the bHLH/PAS superfamily which are transcriptional activators. In mammals, AH receptor nuclear translocator (ARNT), single-minded homologs (SIM1 and SIM2), hypoxia-inducible factor 1 alpha (HIF1A), AH receptor (AHR), neuronal pas domain proteins (NPAS1 and NPAS2), endothelial pas domain protein 1 (EPAS1), mouse ARNT2, and human BMAL1. In *drosophila*, single-minded (SIM), AH receptor nuclear translocator (ARNT), trachealess protein (TRH), and similar protein (SIMA).

- Mammalian transcription factors HES, which repress transcription by acting on two types of DNA sequences, the E box and the N box.

- Mammalian MAD protein (max dimerizer) which acts as transcriptional repressor and may antagonize myc transcriptional activity by competing for max.

- Mammalian Upstream Stimulatory Factor 1 and 2 (USF1 and USF2), which bind to a symmetrical DNA sequence that is found in a variety of viral and cellular promoters.

- Human lyl-1 protein; which is involved, by chromosomal translocation, in T- cell leukemia.

- Human transcription factor AP-4.

- Mouse helix-loop-helix proteins MATH-1 and MATH-2 which activate E box- dependent transcription in collaboration with E47.

- Mammalian stem cell protein (SCL) (also known as tal1), a protein which may play an important role in hemopoietic differentiation. SCL is involved, by chromosomal translocation, in stem-cell leukemia.

- Mammalian proteins Id1 to Id4 [5]. Id (inhibitor of DNA binding) proteins lack a basic DNA-binding domain but are able to form heterodimers with other HLH proteins, thereby inhibiting binding to DNA.

- Drosophila extra-macrochaetae (emc) protein, which participates in sensory organ patterning by antagonizing the neurogenic activity of the achaete- scute complex. Emc is the homolog of mammalian Id proteins.

- Human Sterol Regulatory Element Binding Protein 1 (SREBP-1), a transcriptional activator that binds to the sterol regulatory element 1 (SRE-1) found in the flanking region of the LDLR gene and in other genes.

- Drosophila achaete-scute (AS-C) complex proteins T3 (l'sc), T4 (scute), T5 (achaete) and T8 (asense). The AS-C proteins are involved in the determination of the neuronal precursors in the peripheral nervous system and the central nervous system.

- Mammalian homologs of achaete-scute proteins, the MASH-1 and MASH-2 proteins.

- Drosophila atonal protein (ato) which is involved in neurogenesis.

- Drosophila daughterless (da) protein, which is essential for neurogenesis and sex-determination.

- Drosophila deadpan (dpn), a hairy-like protein involved in the functional differentiation of neurons.

- Drosophila delilah (dei) protein, which is plays an important role in the differentiation of epidermal cells into muscle.

- Drosophila hairy (h) protein, a transcriptional repressor which regulates the embryonic segmentation and adult bristle patterning.

- Drosophila enhancer of split proteins E(spl), that are hairy-like proteins active during neurogenesis. also act as transcriptional repressors.

- Drosophila twist (twi) protein, which is involved in the establishment of germ layers in embryos.

- Maize anthocyanin regulatory proteins R-S and LC.

30

- Yeast phosphate system positive regulatory protein PHO4 which interacts with the upstream activating sequence of several acid phosphatase genes.

- *Neurospora crassa* nuc-1, a protein that activates the transcription of structural genes for phosphorus acquisition.

The schematic representation of the helix-loop-helix domain is shown here:

The signature pattern that had been developed to detect this domain spans completely the second amphipathic helix.

Sequences known to belong to this class detected by the pattern the majority but far from all.
Other sequence(s) detected in SWISS-PROT135.

[2] Garrel J., Campuzano S. BioEssays 13:493-498(1991).

[4] Krause M., Fire A., Harrison S.W., Priess J., Weintraub H. Cell 63:907-919(1990).

[5] Riechmann V., van Cruuchten I., Sablitzky F. *Nucleic Acids Res.* 22:749-755(1994).

Beta-lactamases classes -A, -C, and -D active site

Beta-lactamases (EC 3.5.2.6) [1,2] are enzymes which catalyze the hydrolysis of an amide bond in the beta-lactam ring of antibiotics belonging to the penicillin/cephalosporin family. Four kinds of beta-lactamase have been identified [3]. Class-B enzymes are zinc containing proteins whilst class -A, C and D enzymes are serine hydrolases. The three classes of serine beta-lactamases are evolutionary related and belong to a superfamily [4] that also includes DD-peptidases and a variety of other penicillin-binding proteins (PBP's). All these proteins contain a Ser-x-x-Lys motif, where the serine is the active site residue. Although clearly homologous, the sequences of the three classes of serine beta-lactamases exhibit a large degree of variability and only a small number of residues are conserved in addition to the catalytic serine.

Since a pattern detecting all serine beta-lactamases would also pick up many unrelated sequences, it was decided to provide specific patterns, centered on the active site serine, for each of the three classes.

Consensus pattern [FY]-x-[LIVMFY]-x-S-[TV]-x-K-x(4)-[AGLM]-x(2)-[LC] [S is the active site residue] Sequences known to belong to this class detected by the patternALL class-A beta-lactamases. Other sequence(s) detected in SWISS-PROT7.

Consensus pattern F-E-[LIVM]-G-S-[LIVMG]-[SA]-K [The first S is the active site residue] Sequences known to belong to this class detected by the patternALL class-C beta-lactamases. Other sequence(s) detected in SWISS-PROT NONE.

Consensus pattern [PA]-x-S-[ST]-F-K-[LIV]-[PAL]-x-[STA]-[LI] [S is the active site residue] Sequences known to belong to this class detected by the patternALL class-D beta-lactamases. Other sequence(s) detected in SWISS-PROT NONE.

[1] Ambler R.P. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 289:321-331(1980).

[2] Pastor N., Pinero D., Valdes A.M., Soberon X. Mol. Microbiol. 4:1957-1965(1990).

[3] Bush K. Antimicrob. Agents Chemother. 33:259-263(1989).

[4] Joris B., Ghuyssen J.-M., Dive G., Renard A., Dideberg O., Charlier P., Frere J.M., Kelly J.A., Boyington J.C., Moews P.C., Knox J.R. Biochem. J. 250:313-324(1988).

920. Biotin protein ligase (BPL)

Biotin is covalently attached at the active site of certain enzymes that transfer carbon dioxide from bicarbonate to organic acids to form cellular metabolites. Biotin protein ligase (BPL) is the enzyme responsible for attaching biotin to a specific lysine at the active site of biotin enzymes. Each organism probably has only one BPL. Biotin attachment is a two step reaction that results in the formation of an amide linkage between the carboxyl group of biotin and the epsilon-amino group of the modified lysine [2].

Number of members: 26

[1] Wilson KP, Shewchuk LM, Brennan RG, Otsuka AJ, Matthews BW; Medline: 93028443 "Escherichia coli biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin- and DNA-binding domains." Proc Natl Acad Sci USA 1992;89:9257-9261.

[2] Chapman-Smith A, Cronan JE Jr; Medline: 10470036 "The enzymatic biotinylation of proteins: a post-translational modification of exceptional specificity." Trends Biochem Sci 1999;24:359-363.

921. (BRCA2_repeat)

The alignment covers only the most conserved region of the repeat. Respiratory-chain NADH dehydrogenase 30 Kd subunit signature

[1] Bork P, Blomberg N, Nilges M; Medline: 96241568 "Internal repeats in the BRCA2 protein sequence." Nat Genet 1996;13:22-23.

Number of members: 63

922. (C6)

This domain of unknown function is found in the C. elegans protein Swiss:Q19522. It is presumed to be an extracellular domain. The C6 domain contains six conserved cysteine

756

residues in most copies of the domain. However some copies of the domain are missing cysteine residues 1 and 3 suggesting that these form a disulphide bridge.

Number of members: 23

5 923. Cadherin cytoplasmic region (Cadherin_C_term)

Cadherins are vital in cell-cell adhesion during tissue differentiation. Cadherins are linked to the cytoskeleton by catenins. Catenins bind to the cytoplasmic tail of the cadherin. Cadherins cluster to form foci of homophilic binding units. A key determinant to the strength of the binding that it is mediated by cadherins is the juxtamembrane region of the cadherin. This region induces clustering and also binds to the protein p120ctn [1].

Number of members: 59

[1] Yap AS, Niessen CM, Gumbiner BM; Medline: 98234411 "The juxtamembrane region of the cadherin cytoplasmic tail supports lateral clustering, adhesive strengthening, and interaction with p120ctn." J Cell Biol 1998;141:779-789.

[2] Barth AI, Nathke IS, Nelson WJ; Medline: 97471931 "Cadherins, catenins and APC protein: interplay between cytoskeletal complexes and signaling pathways." Curr Opin Cell Biol 1997;9:683-690.

[3] Braga VM, Machesky LM, Hall A, Hotchin NA; Medline: 97327766 "The small GTPases Rho and Rac are required for the establishment of cadherin-dependent cell-cell contacts." J Cell Biol 1997;137:1421-1431.

25 924. Clathrin propeller repeat (Clathrin_propel)

Clathrin is the scaffold protein of the basket-like coat that surrounds coated vesicles. The soluble assembly unit, a triskelion, contains three heavy chains and three light chains in an extended three-legged structure. Each leg contains one heavy and one light chain. The N-terminus of the heavy chain is known as the globular domain, and is composed of seven repeats which form a beta propeller [1].

Number of members: 61

[1] ter Haar E, Musacchio A, Harrison SC, Kirchhausen T; Medline: 99043510 "Atomic structure of clathrin: a beta propeller terminal domain joins an alpha zigzag linker." Cell. 1998;95:563-573.

5 925. Respiratory-chain NADH dehydrogenase 30 Kd subunit signature (complex1_30Kd)

Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex there is one with a molecular weight of 30 Kd (in mammals) which has been found to be:

- Nuclear encoded, as a precursor form with a transit peptide in mammals, and in *Neurospora crassa*.

- Mitochondrial encoded in *Paramecium* (protein P1), and in the slime mold *Dictyostelium discoideum* (ORF 209).

- Chloroplast encoded in various higher plants (ORF 159). It is also present in bacteria:

- In the cyanobacteria *Synechocystis* strain PCC 6803 (gene *ndhJ*).

- Subunit C of *Escherichia coli* NADH-ubiquinone oxidoreductase (gene *nuoC*).

- Subunit NQO5 of *Paracoccus denitrificans* NADH-ubiquinone oxidoreductase.

This protein, in its mature form, consists of from 157 to 266 amino acid residues. The best conserved region is located in the C-terminal section and can be used as a signature pattern.

25 Consensus pattern E-R-E-x(2)-[DE]-[LIVMFY](2)-x(6)-[HK]-x(3)-[KRP]-x-[LIVM]-[LIVMYS] Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROT/NONE.

[1] Ragan C.I. Curr. Top. Bioenerg. 15:1-36(1987).

30 [2]Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).

926. Respiratory-chain NADH dehydrogenase 49 Kd subunit signature (complex1_49Kd)

758

Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex there is one with a molecular weight of 49 Kd (in mammals), which is the third largest subunit of complex I and is a component of the iron-sulfur (IP) fragment of the enzyme. It seems to bind a 4Fe-4S iron-sulfur cluster. The 49 Kd subunit has been found to be:

- Nuclear encoded, as a precursor form with a transit peptide in mammals, and in *Neurospora crassa*.
- Mitochondrial encoded in protozoan such as *Paramecium* (ORF 400), *Leishmania* and *Trypanosoma* (MURF 3).
- Chloroplast encoded in various higher plants (ORF 392).

The 49 Kd subunit is highly similar to [3,4]:

- Subunit D of *Escherichia coli* NADH-ubiquinone oxidoreductase (gene *nuoD*).
- Subunit NQO4 of *Paracoccus denitrificans* NADH-ubiquinone oxidoreductase.
- Subunit 5 of *Escherichia coli* formate hydrogenlyase (gene *hycE*).
- Subunit G of *Escherichia coli* hydrogenase-4 (gene *hyfG*).

A highly conserved region was selected as signature pattern, located in the N-terminal section of this subunit.

Consensus pattern [LIVMH]-H-[RT]-[GA]-x-E-K-[LIVMTN]-x-E-x-[KRQ] Sequences known to belong to this class detected by the patternALL.

- [1] Ragan C.I. Curr. Top. Bioenerg. 15:1-36(1987).
- [2] Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).
- [3] Fearnley I.M., Walker J.E. Biochim. Biophys. Acta 1140:105-134(1992).
- [4] Weidner U., Geier S., Ptöck A., Friedrich T., Leif H., Weiss H. J. Mol. Biol. 233:109-122(1993).

927. (COX2)

Cytochrome c oxidase (EC 1.9.3.1) [1,2] is an oligomeric enzymatic complex which is a component of the respiratory chain and is involved in the transfer of electrons from cytochrome c to oxygen. In eukaryotes this enzyme complex is located in the mitochondrial inner membrane; in aerobic prokaryotes it is found in the plasma membrane. The enzyme complex consists of 3-4 subunits (prokaryotes) to up to 13 polypeptides (mammals).

Subunit 2 (CO II) transfers the electrons from cytochrome c to the catalytic subunit 1. It contains two adjacent transmembrane regions in its N-terminus and the major part of the protein is exposed to the periplasmic or to the mitochondrial intermembrane space, respectively. CO II provides the substrate-binding site and contains a copper center called Cu(A), probably the primary acceptor in cytochrome c oxidase. An exception is the corresponding subunit of the cbb3-type oxidase which lacks the copper A redox-center. Several bacterial CO II have a C-terminal extension that contains a covalently bound heme c.

It has been shown [3,4] that nitrous oxide reductase (EC 1.7.99.6) (gene nosZ) of *Pseudomonas* has sequence similarity in its C-terminus to CO II. This enzyme is part of the bacterial respiratory system which is activated under anaerobic conditions in the presence of nitrate or nitrous oxide. NosZ is a periplasmic homodimer that contains a dinuclear copper center, probably located in a 3-dimensional fold similar to the cupredoxin-like fold that has been suggested for the copper-binding site of CO II [3].

The dinuclear purple copper center is formed by 2 histidines and 2 cysteines [5]. This region was used as a signature pattern. The conserved valine and the conserved methionine are said to be involved in stabilizing the copper-binding fold by interacting with each other.

Consensus pattern V-x-H-x(33,40)-C-x(3)-C-x(3)-H-x(2)-M [The two C's and two H's are copper ligands] Sequences known to belong to this class detected by the patternALL, except for *Paramecium primaurelia* as well as in some plants where the pattern ends with Thr; an RNA editing event at this position could change this Thr to Met.

Note: cytochrome cbb(3) subunit 2 does not belong to this family.

- [1] Capaldi R.A., Malatesta F., Darley-Usmar V.M. *Biochim. Biophys. Acta* 726:135-148(1983).
- [2] Garcia-Horsman J.A., Barquera B., Rumbley J., Ma J., Gennis R.B. *J. Bacteriol.* 176:5587-5600(1994).
- [3] van der Oost J., Lappalainen P., Musacchio A., Warne A., Lemieux L., Rumbley J., Gennis R.B., Aasa R., Pascher T., Malmstrom B.G., Saraste M. *EMBO J.* 11:3209-3217(1992).
- [4] Zumft W.G., Dreutsch A., Loechele S., Cuypers H., Friedrich B., Schneider B. *Eur. J. Biochem.* 208:31-40(1992).

928. Cytochrome C assembly protein (CytC_asm)

This family consists of various proteins involved in cytochrome c assembly from mitochondria and bacteria; CycK from *Rhizobium*[3], CcmC from *E. coli* and *Paracoccus denitrificans* [2,1] and orf240 from wheat mitochondria [4]. The members of this family are probably integral membrane proteins with six predicted transmembrane helices. It has been proposed that members of this family comprise a membrane component of an ABC (ATP binding cassette) transporter complex. It is also proposed that this transporter is necessary for transport of some component needed for cytochrome c assembly. One member CycK contains a putative heme-binding motif [3], orf240 also contains a putative heme-binding motif and is a proposed ABC transporter with c-type heme as its proposed substrate [4]. However it seems unlikely that all members of this family transport heme nor c-type apocytochromes because CcmC in the putative CcmABC transporter transports neither [1].

Number of members: 67

- [1] Page D, Pearce DA, Norris HA, Ferguson SJ; Medline: 97195802 "The *Paracoccus denitrificans* ccmA, B and C genes: cloning and sequencing, and analysis of the potential of their products to form a haem or apo-c-type cytochrome transporter. *MICROBIOLOGY* 1997;143:563-576.
- [2] Thoeny-meyer L, Fischer F, Kunzler P, Ritz D, Hennecke H; Medline: 95362656 "Escherichia coli genes required for cytochrome c maturation." *J. BACTERIOL* 1995;177:4321-4326.

[3] Delgado MJ, Yeoman KH, Wu G, Vargas C, Davies A, Poole RK, Johnston AWB, Downie JA; Medline: 95394794 "Characterization of the *cycHJKL* genes involved in cytochrome c biogenesis and symbiotic nitrogen fixation in *Rhizobium leguminosarum*." J. BACTERIOL 1995;177:4927-4934.

- 5 [4] Bonnard G, Grienemberger JM; Medline: 95124303 "A gene proposed to encode a transmembrane domain of an ABC transporter is expressed in wheat mitochondria." MOL. GEN. GENET 1995;246:91-99.

929. Cytochrome b559 subunits heme-binding site signature (cytochr_b559)

10

Cytochrome b559 [1] is an essential component of photosystem II complex from oxygenic photosynthetic organisms. It is an integral thylakoid membrane protein composed of two subunits, alpha (gene *psbE*) and beta (gene *psbF*), each of which contains a histidine residue located in a transmembrane region. The two histidines coordinate the heme iron of cytochrome b559.

15

The region around the heme-binding residue of both subunits is very similar and can be used as a signature pattern.

20

Consensus pattern[LIV]-x-[ST]-[LIVF]-R-[FYW]-x(2)-[IV]-H-[STGA]-[LIV]- [STGA]-[IV]-P [H is the heme iron ligand] Sequences known to belong to this class detected by the patternALL. Other sequence(s) detected in SWISS-PROT NONE.

[1] Pakrasi H.B., de Ciechi P., Whitmarsh J. EMBO J. 10:1619-1627(1991).

25

930. Cytochrome b/b6 signatures (Cytochrome_b)

30

In the mitochondrion of eukaryotes and in aerobic prokaryotes, cytochrome b is a component of respiratory chain complex III (EC 1.10.2.2) - also known as the bc1 complex or ubiquinol-cytochrome c reductase. In plant chloroplasts and cyanobacteria, there is a analogous protein, cytochrome b6, a component of the plastoquinone-plastocyanin reductase (EC 1.10.99.1), also known as the b6f complex.

Cytochrome b/b6 [1,2] is an integral membrane protein of approximately 400 amino acid residues that probably has 8 transmembrane segments. In plants and cyanobacteria, cytochrome b6 consists of two subunits encoded by the petB and petD genes. The sequence of petB is colinear with the N-terminal part of mitochondrial cytochrome b, while petD corresponds to the C-terminal part. Cytochrome b/b6 non-covalently binds two heme groups, known as b562 and b566. Four conserved histidine residues are postulated to be the ligands of the iron atoms of these two heme groups.

Apart from regions around some of the histidine heme ligands, there are a few conserved regions in the sequence of b/b6. The best conserved of these regions includes an invariant P-E-W triplet which lies in the loop that separates the fifth and sixth transmembrane segments. It seems to be important for electron transfer at the ubiquinone redox site - called Qz or Qo (where o stands for outside) - located on the outer side of the membrane.

A schematic representation of the structure of cytochrome b/b6 is shown below.

```

+---Fe-b562---+ | +---Fe-b566--|-+ |||
xxxxxxxxxxxxHxHxxxxxxxxxxxxHxHxxxxxxxxxxPEWxxxxxxxxxxxxxxxxxxxx <-----
---Cytochrome-b-----> <---Cytochrome-b6-petB-----><---Cytochrome-
b6-petD----->

```

Two signature patterns were developed for cytochrome b/b6. The first includes the first conserved histidine of b/b6, which is a heme b562 ligand; the second includes the conserved PEW triplet.

Consensus pattern [DENQ]-x(3)-G-[FYWMQ]-x-[LIVMF]-R-x(2)-H [H is a heme b562 ligand] Sequences known to belong to this class detected by the patternALL, except for 5 sequences.

Consensus pattern P-[DE]-W-[FY]-[LFY](2) Sequences known to belong to this class detected by the patternALL, except for *Odocoileus hemionus* (mule deer) and *Paramecium tetraurelia* cytochrome b.

[1] Howell N. J. Mol. Evol. 29:157-169(1989).

[2] Esposti M.D., de Vries S., Crimi M., Ghelli A., Patarnello T., Meyer A. Biochim. Biophys. Acta 1143:243-271(1993).

5

931. Phorbol esters / diacylglycerol binding domain (DAG_PE-bind)

Diacylglycerol (DAG) is an important second messenger. Phorbol esters (PE) are analogues of DAG and potent tumor promoters that cause a variety of physiological changes when administered to both cells and tissues. DAG activates a family of serine/threonine protein kinases, collectively known as protein kinase C (PKC) [1]. Phorbol esters can directly stimulate PKC. The N- terminal region of PKC, known as C1, has been shown [2] to bind PE and DAG in a phospholipid and zinc-dependent fashion. The C1 region contains one or two copies (depending on the isozyme of PKC) of a cysteine-rich domain about 50 amino-acid residues long and essential for DAG/PE-binding. Such a domain has also been found in the following proteins:

10

15

20

25

30

- Diacylglycerol kinase (EC 2.7.1.107) (DGK) [3], the enzyme that converts DAG into phosphatidate. It contains two copies of the DAG/PE-binding domain in its N-terminal section. At least five different forms of DGK are known in mammals.
- N-chimaerin. A brain specific protein which shows sequence similarities with the BCR protein at its C-terminal part and contains a single copy of the DAG/PE-binding domain at its N-terminal part. It has been shown [4,5] to be able to bind phorbol esters.
- The raf/mil family of serine/threonine protein kinases. These protein kinases contain a single N-terminal copy of the DAG/PE-binding domain.
- The unc-13 protein from *Caenorhabditis elegans*. Its function is not known but it contains a copy of the DAG/PE-binding domain in its central section and has been shown to bind specifically to a phorbol ester in the presence of calcium [6].
- The vav oncogene. Vav was generated by a genetic rearrangement during gene transfer assays. Its expression seems to be restricted to cells of hematopoietic origin. Vav seems [5,7] to contain a DAG/PE-binding domain in the central part of the protein.
- The *Drosophila* GTPase activating protein rotund.

The DAG/PE-binding domain binds two zinc ions; the ligands of these metal ions are probably the six cysteines and two histidines that are conserved in this domain. A signature pattern was developed that spans completely the DAG/PE domain.

- 5 Consensus pattern H-x-[LIVMFYW]-x(8,11)-C-x(2)-C-x(3)-[LIVMFC]-x(5,10)- C-x(2)-C-x(4)-[HD]-x(2)-C-x(5,9)-C [All the C and H are involved in binding Zinc] Sequences known to belong to this class detected by the pattern ALL, except a few DGK's.

[1] Azzi A., Boscoboinik D., Hensey C. Eur. J. Biochem. 208:547-557(1992).

- 10 [2] Ono Y., Fujii T., Igarashi K., Kuno T., Tanaka C, Kikkawa U., Nishizuka Y. Proc. Natl. Acad. Sci. U.S.A. 86:4868-4871(1989).

[3] Sakane F., Yamada K., Kanoh H., Yokoyama C., Tanabe T. Nature 344:345-348(1990).

[4] Ahmed S., Kozma R., Monfries C., Hall C., Lim H.H., Smith P., Lim L. Biochem. J. 272:767-773(1990).

15 [5] Ahmed S., Kozma R., Lee J., Monfries C., Harden N., Lim L. Biochem. J. 280:233-241(1991).

[6] Ahmed S., Maruyama I.N., Kozma R., Lee J., Brenner S., Lim L. Biochem. J. 287:995-999(1992).

[7] Boguski M.S., Bairoch A., Attwood T.K., Michaels G.S. Nature 358:113-113(1992).

20 932. 3-dehydroquinate synthase (DHQ_synthase)

[1] Barten R, Meyer TF; Medline: 98273626 "Cloning and characterisation of the *Neisseria gonorrhoeae* *aroB* gene." Mol Gen Genet 1998;258:34-44.

- 25 [2] Hawkins AR, Lamb HK; Medline: 96048023 "The molecular biology of multidomain proteins. Selected examples." Eur J Biochem 1995;232:7-18.

The 3-dehydroquinate synthase EC:4.6.1.3 domain is present in isolation in various bacterial 3-dehydroquinate synthases and also present as a domain in the pentafunctional AROM polypeptide Swiss:P07547 [2]. 3-dehydroquinate (DHQ) synthase catalyses the formation of dehydroquinate (DHQ) and orthophosphate from 3-deoxy-D-arabino heptulosonic 7 phosphate [1]. This reaction is part of the shikimate pathway which is involved in the biosynthesis of aromatic amino acids.

30

Number of members: 25

933. Dihydrofolate reductase signature (DiHfolate_red)

- 5 Dihydrofolate reductases (EC 1.5.1.3) [1] are ubiquitous enzymes which catalyze the reduction of folic acid into tetrahydrofolic acid. They can be inhibited by a number of antagonists such as trimethoprim and methotrexate which are used as antibacterial or anticancerous agents. A signature pattern was derived from a region in the N-terminal part of these enzymes, which includes a conserved Pro-Trp dipeptide; the tryptophan has been
10 shown [2] to be involved in the binding of substrate by the enzyme.

Consensus pattern[LVAGC]-[LIF]-G-x(4)-[LIVMF]-P-W-x(4,5)-[DE]-x(3)-[FYIV]-
x(3)-[STIQ] Sequences known to belong to this class detected by the patternALL, except for
15 type II bacterial, plasmid-encoded, dihydrofolate reductases which do not belong to the same class of enzymes.

[1] Harpers' Review of Biochemistry, Lange, Los Altos (1985).

[2] Bolin J.T., Filman D.J., Matthews D.A., Hamlin R.C., Kraut J. J. Biol. Chem. 257:13650-
13662(1982).

20 934. (DIL)

[1] Ponting CP; Medline: 95397417 "AF-6/cno: neither a kinesin nor a myosin, but a bit of both." Trends Biochem Sci 1995;20:265-266.

25 Number of members: 31

935. (DNA_gyraseB_C)

DNA topoisomerase II signature (cross-reference = TOPOISOMERASE_II)

- 30 DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type II topoisomerases are ATP-dependent and act by passing a DNA segment through a transient double-strand break.

Topoisomerase II is found in phages, archaebacteria, prokaryotes, eukaryotes, and in African Swine Fever virus (ASF). In bacteriophage T4 topoisomerase II consists of three subunits (the product of genes 39, 52 and 60). In prokaryotes and in archaebacteria the enzyme, known as DNA gyrase, consists of two subunits (genes *gyrA* and *gyrB* [E2]). In some bacteria, a second type II topoisomerase has been identified; it is known as topoisomerase IV and is required for chromosome segregation, it also consists of two subunits (genes *parC* and *parE*). In eukaryotes, type II topoisomerase is a homodimer.

There are many regions of sequence homology between the different subtypes of topoisomerase II. The relation between the different subunits is shown in the following representation:

```

<-----About-1400-residues----->
[-----Protein 39-*-----][----Protein 52----] Phage T4
[-----gyrB-----*-----][-----gyrA-----] Prokaryote II
Archaebacteria
[-----parE-----*-----][-----parD-----] Prokaryote IV
[-----*-----] Eukaryote and ASF
'!': Position of the pattern.

```

As a signature pattern for this family of proteins, a region was selected that contains a highly conserved pentapeptide. The pattern is located in *gyrB*, in *parE*, and in protein 39 of phage T4 topoisomerase.

Consensus pattern [LIVMA]-x-E-G-[DN]-S-A-x-[STAG] Sequences known to belong to this class detected by the pattern ALL.

[1] Sternglanz R. Curr. Opin. Cell Biol. 1:533-535(1990).

[2] Bjornsti M.-A. Curr. Opin. Struct. Biol. 1:99-103(1991).

[3] Sharma A., Mondragon A. Curr. Opin. Struct. Biol. 5:39-47(1995).

[4] Roca J. Trends Biochem. Sci. 20:156-160(1995).

936. (DNA_topoisolIV)

DNA topoisomerase II signature (cross-reference = TOPOISOMERASE_II)

DNA topoisomerase I (EC 5.99.1.2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type II topoisomerases are ATP-dependent and act by passing a DNA segment through a transient double-strand break. Topoisomerase II is found in phages, archaebacteria, prokaryotes, eukaryotes, and in African Swine Fever virus (ASF). In bacteriophage T4 topoisomerase II consists of three subunits (the product of genes 39, 52 and 60). In prokaryotes and in archaebacteria the enzyme, known as DNA gyrase, consists of two subunits (genes *gyrA* and *gyrB* [E2]). In some bacteria, a second type II topoisomerase has been identified; it is known as topoisomerase IV and is required for chromosome segregation, it also consists of two subunits (genes *parC* and *parE*). In eukaryotes, type II topoisomerase is a homodimer.

There are many regions of sequence homology between the different subtypes of topoisomerase II. The relation between the different subunits is shown in the following representation:

```
<-----About-1400-residues----->
[-----Protein 39-*-----][----Protein 52----] Phage T4
[-----gyrB-----*-----][-----gyrA-----] Prokaryote II Archaebacteria
[-----parE-----*-----][-----parD-----] Prokaryote IV
[-----*-----] Eukaryote and ASF
```

'*': Position of the pattern.

As a signature pattern for this family of proteins, a region was selected that contains a highly conserved pentapeptide. The pattern is located in *gyrB*, in *parE*, and in protein 39 of phage T4 topoisomerase.

Consensus pattern [LIVMA]-x-E-G-[DN]-S-A-x-[STAG] Sequences known to belong to this class detected by the patternALL.

[1] Sternglanz R. Curr. Opin. Cell Biol. 1:533-535(1990).

[2] Bjornsti M.-A. Curr. Opin. Struct. Biol. 1:99-103(1991).

[3] Sharma A., Mondragon A. Curr. Opin. Struct. Biol. 5:39-47(1995).

[4] Roca J. Trends Biochem. Sci. 20:156-160(1995).

937. Prolyl oligopeptidase family serine active site (DPPIV_N_term)

5

The prolyl oligopeptidase family [1,2,3] consist of a number of evolutionary related peptidases whose catalytic activity seems to be provided by a charge relay system similar to that of the trypsin family of serine proteases, but which evolved by independent convergent evolution. The known members of this family are listed below.

- 10 - Prolyl endopeptidase (EC 3.4.21.26) (PE) (also called post-proline cleaving enzyme). PE is an enzyme that cleaves peptide bonds on the C-terminal side of prolyl residues. The sequence of PE has been obtained from a mammalian species (pig) and from bacteria (*Flavobacterium meningosepticum* and *Aeromonas hydrophila*); there is a high degree of sequence conservation between these sequences.
- 15 - *Escherichia coli* protease II (EC 3.4.21.83) (oligopeptidase B) (gene prtB) which cleaves peptide bonds on the C-terminal side of lysyl and arginyl residues.
- Dipeptidyl peptidase IV (EC 3.4.14.5) (DPP IV). DPP IV is an enzyme that removes N-terminal dipeptides sequentially from polypeptides having unsubstituted N-termini provided that the penultimate residue is proline.
- 20 - Yeast vacuolar dipeptidyl aminopeptidase A (DPAP A) (gene: STE13) which is responsible for the proteolytic maturation of the alpha-factor precursor.
- Yeast vacuolar dipeptidyl aminopeptidase B (DPAP B) (gene: DAP2).
- Acylamino-acid-releasing enzyme (EC 3.4.19.1) (acyl-peptide hydrolase). This enzyme catalyzes the hydrolysis of the amino-terminal peptide bond of an N-acetylated protein to
- 25 generate a N-acetylated amino acid and a protein with a free amino-terminus.

A conserved serine residue has experimentally been shown (in *E.coli* protease II as well as in pig and bacterial PE) to be necessary for the catalytic mechanism. This serine, which is part of the catalytic triad (Ser, His, Asp), is generally located about 150 residues away from the C-terminal extremity of these enzymes (which are all proteins that contains about 700 to 800 amino acids).

30

769

Consensus pattern D-x(3)-A-x(3)-[LIVMFYW]-x(14)-G-x-S-x-G-G-[LIVMFYW](2) [S is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for yeast DPAP A.

5 Note: these proteins belong to families S9A/S9B/S9C in the classification of peptidases [4,E1].

[1] Rawlings N.D., Polgar L., Barrett A.J. Biochem. J. 279:907-911(1991).

[2] Barrett A.J., Rawlings N.D. Biol. Chem. Hoppe-Seyler 373:353-360(1992).

10 [3] Polgar L., Szabo E. Biol. Chem. Hoppe-Seyler 373:361-366(1992).

[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).

938. Deoxyhypusine synthase (DS)

15 Eukaryotic initiation factor 5A (eIF-5A) contains an unusual amino acid, hypusine [N epsilon-(4-aminobutyl-2-hydroxy)lysine]. The first step in the post-translational formation of hypusine is catalysed by the enzyme deoxyhypusine synthase (DS) EC:1.1.1.249. The modified version of eIF-5A, and DS, are required for eukaryotic cell proliferation [1].

20 Number of members: 9

[1] Liao DI, Wolff EC, Park MH, Davies DR; Medline: 98154315 "Crystal structure of the NAD complex of human deoxyhypusine synthase: an enzyme with a ball-and-chain mechanism for blocking the active site." Structure 1998;6:23-32.

939. (DUF21)

30 Many of the sequences in this family are annotated as hemolysins, however this is due to a similarity to Swiss:Q54318 that does not contain this domain. This domain is found in the N-terminus of the proteins adjacent to two intracellular CBS domains CBS.

Number of members: 42

940. (DUF59)

This family includes prokaryotic proteins of unknown function. The family also includes
5 PhaH Swiss:O84984 from *Pseudomonas putida*. PhaH forms a complex with PhaF
Swiss:O84982, PhaG Swiss:O84983 and PhaI Swiss:O84985, which hydroxylates
phenylacetic acid to 2-hydroxyphenylacetic acid [1]. So members of this family may all be
components of ring hydroxylating complexes.

Number of members: 15

[1] Olivera ER, Minambres B, Garcia B, Muniz C, Moreno MA, Ferrandez A, Diaz E, Garcia
10 JL, Luengo JM; Medline: 98263372 "Molecular characterization of the phenylacetic acid
catabolic pathway in *Pseudomonas putida* U: the phenylacetyl-CoA catabolon." Proc Natl
Acad Sci U S A 1998;95:6419-6424.

941. (DUF82)

The protein contains four conserved cysteines that may be involved in metal binding or
disulphide bridges.

Number of members: 4

942. Riboflavin kinase / FAD synthetase (FAD_Synth)

This family consists part of the bifunctional enzyme riboflavin kinase / FAD synthetase.

25 These enzymes have both ATP:riboflavin 5'-phospho transferase and ATP:FMN-
adenylyltransferase activities [1]. They catalyse the 5'-phosphorylation of riboflavin to FMN
and the adenylation of FMN to FAD [1].

CAUTION: It is not clear if this region of the enzymes catalyses either or both of the
enzymatic reactions.

30 Number of members: 27

[1] Manstein DJ, Pai EF; Medline: 87057286 "Purification and characterization of FAD
synthetase from *Brevibacterium ammoniagenes*." J Biol Chem 1986;261:16169-16173.

943. [2Fe-2S] binding domain (fer2_2)

[1] Romao MJ, Archer M, Moura I, Moura JJ, LeGall J, Engh R, Schneider M, Hof P, Huber R; Medline: 96072968 "Crystal structure of the xanthine oxidase-related aldehyde oxidoreductase from *D. gigas*." *Science* 1995;270:1170-1176.

Number of members: 53

944. Filovirus glycoprotein (Filo_glycop)

This family includes an extracellular region from the envelope glycoprotein of Ebola and Marburg viruses. This region is also produced as a separate transcript that gives rise to a non-structural, secreted glycoprotein, which is produced in large amounts and has an unknown function [1]. Processing of this protein may be involved in viral pathogenicity [2].

Number of members: 23

[1] Volchkov VE, Feldmann H, Volchkova VA, Klenk HD; Medline: 98245155 "Processing of the Ebola virus glycoprotein by the proprotein convertase furin." *Proc Natl Acad Sci U S A* 1998;95:5762-5767.

[2] Sanchez A, Trappier SG, Mahy BW, Peters CJ, Nichol ST; Medline: 96195018 "The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing." *Proc Natl Acad Sci U S A* 1996;93:3602-3607.

945. Frataxin-like domain (Frataxin_Cyay)

This family contains proteins that have a domain related to the globular C-terminus of Frataxin the protein that is mutated in Friedreich's ataxia. This domain is found in a family of bacterial proteins. The function of this domain is currently unknown.

Number of members: 12

[1] Gibson TJ, Koonin EV, Musco G, Pastore A, Bork P; Medline: 97084946 "Friedreich's ataxia protein: phylogenetic evidence for mitochondrial dysfunction." *Trends Neurosci* 1996;19:465-468.

946. (GAF)

Domain present in phytochromes and cGMP-specific phosphodiesterases.

5 Number of members: 296

[1] Aravind L, Ponting CP; Medline: 98094688 "The GAF domain: an evolutionary link between diverse phototransducing proteins." Trends Biochem Sci 1997;22:458-459.

10 947. Galaptin signature (Gal-bind_lectin)

All vertebrates synthesize soluble galactoside-binding lectins [1,2,3] (also known as galectins, galaptins or S-lectin). These carbohydrate-binding proteins are developmentally regulated. Although their exact physiological role is not yet clear they seem to be involved in differentiation, cellular regulation and tissue construction. The sequence of galactoside-binding lectins from electric eel (electrolectin), conger eel (congerin), chicken and a number of mammalian species is known. These lectins are proteins of about 130 to 140 amino acid residues (14 Kd to 16 Kd).

A number of other proteins are known to belong to this family:

- Galectin-3 (also known as MAC-2 antigen; CBP-35 or IgE-binding protein), a 35 Kd lectin which binds immunoglobulin E and which is composed of two domains: a N-terminal domain that consist of tandem repeats of a glycine/ proline-rich sequence and a C-terminal galaptin domain.

- Galectin-4 [4], which is composed of two galaptin domains.

- Galectin-5.

- Galectin-7 [5], a keratinocyte protein which could be involved in cell-cell and/or cell-matrix interactions necessary for normal growth control.

- Galectin-8 [6], which is composed of two galaptin domains.

- Galectin-9 [7], which is composed of two galaptin domains.

- Human eosinophil lysophospholipase (EC 3.1.1.5) [8] (Charcot-Leyden crystal protein), a protein that may have both an enzymatic and a lectin activities. It forms hexagonal

bipyramidal crystals in tissues and secretions from sites of eosinophil-associated inflammation.

- *Caenorhabditis elegans* 32 Kd lactose-binding lectin [9]. This lectin is composed of two galactin domains.

5 - *Caenorhabditis elegans* lec-7 and lec-8.

One of the conserved regions of these lectins contains a tryptophan that has been shown [10] to be essential to the binding of galactosides. This region was used as a signature pattern for these proteins.

10 Consensus pattern W-[GEK]-x-[EQ]-x-[KRE]-x(3,6)-[PCTF]-[LIVMF]-[NQEKGSKV]-x-[GH]-x(3)-[DENKHS]-[LIVMFC] [W binds carbohydrate] Sequences known to belong to this class detected by the pattern ALL, except for pig galectin 4.

[1] Barondes S.H., Gitt M.A., Leffler H., Cooper D.N.W. *Biochimie* 70:1627-1632(1988).

15 [2] Hirabayashi J., Kasai K.-I. *J. Biochem.* 104:1-4(1988).

[3] Barondes S.H., Castronovo V., Cooper D.N.W., Cummings R.D., Drickamer K., Feizi T., Gitt M.A., Hirabayashi J., Hughes C., Kasai K.-I., Leffler H., Liu F.-T., Lotan R., Mercurio A.M., Monsigny M., Pillair S., Poirer F., Raz A., Rigby P.W.J., Rini J.M., Wang J.L. *Cell* 76:597-598(1994).

20 [4] Oda Y., Herrmann J., Gitt M., Turck C.W., Burlingame A.L., Barondes S.H., Leffler H. *J. Biol. Chem.* 268:5929-5939(1993).

[5] Madsen P., Rasmussen H.H., Flint T., Gromov P., Kruse T.A., Honore B., Vorum H., Celis J.E. *J. Biol. Chem.* 270:5823-5829(1995).

[6] Hadari Y.R., Paz K., Dekel R., Mestrovic T., Accili D., Zick Y. *J. Biol. Chem.* 270:3447-3453(1995).

25 [7] Wada J., Kanwar Y.S. *J. Biol. Chem.* 272:6078-6086(1997).

[8] Ackerman S.J., Corrette S.E., Rosenberg H.F., Bennett J.C., Mastrianni D.M., Nicholson-Weller A., Weller P.F., Chin D.T., Tenen D.G. *J. Immunol.* 150:456-468(1993).

[9] Hirabayashi J., Satoh M., Kasai K.-I. *J. Biol. Chem.* 267:15485-15490(1992).

30 [10] Abbott W.M., Feizi T. *J. Biol. Chem.* 266:5552-5557(1991).

948. (GARS) Phosphoribosylglycinamide synthetase signature (phosphoribosylamine glycine ligase)

PROSITE: PDOC00164; cross-reference(s): PS00184

[1] catalyzes the second step in the de novo biosynthesis of purine, the ATP-dependent addition of 5-phosphoribosylamine to glycine to form 5'phosphoribosylglycinamide.

5 In bacteria GARS is a monofunctional enzyme (encoded by the purD gene), in of a bifunctional enzyme (encoded by the ADE5,7 gene), in higher eukaryotes it is part, with AIRS and with phosphoribosylglycinamide formyltransferase (GART) of a trifunctional enzyme (GARS-AIRS-GART).

10 The sequence of GARS is well conserved. A highly conserved octapeptide was selected as a signature pattern.

Consensus patternR-F-G-D-P-E-x-[QM]

Sequences known to belong to this class detected by the patternALL.

15 [1]Aiba A., Mizobuchi K. J. Biol. Chem. 264:21239-21246(1989).

949. GLTT - GLTT repeat (12 copies)

This short repeat of unknown function is found in multiple copies in several C. elegans proteins. The repeat is five residues long and consists of XGLTT where X can be any amino acid. Number of members: 34.

950. Glu_synthase - Conserved region in glutamate synthase

25 This family represents a region of the glutamate synthase protein. This region is expressed as a separate subunit in the glutamate synthase alpha subunit from archaebacteria, or part of a large multidomain enzyme in other organisms. The aligned region of these proteins contains a putative FMN binding site and Fe-S cluster. Number of members: 44.

30 [1] Medline: 97082505. Sequence of the GLT1 gene from Saccharomyces cerevisiae reveals the domain structure of yeast glutamate synthase. Filetici P, Martegani MP, Valenzuela L, Gonzalez A, Ballario P; Yeast 1996;12:1359-1366.

951. (Glyco_hydro_2) Glycosyl hydrolases family 2 signatures

GLYCOSYL_HYDROL_F2_1; PS00608; GLYCOSYL_HYDROL_F2_2

It has been shown [1,2,E1] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:

-Beta-galactosidases (EC 3.2.1.23) from bacteria such as *Escherichia coli* (genes *lacZ* and *ebgA*), *Clostridium acetobutylicum*, *Clostridium thermosulfurogenes*, *Klebsiella pneumoniae*, *Lactobacillus delbrueckii*, or *Streptococcus thermophilus* and from the fungi *Kluyveromyces lactis*.

-Beta-glucuronidase (EC 3.2.1.31) from *Escherichia coli* (gene *uidA*) and from mammals.

One of the conserved regions in these enzymes is centered on a conserved glutamic acid residue which has been shown [3], in *Escherichia coli lacZ*, to be the general acid/base catalyst in the active site of the enzyme. This region has been used as a signature pattern. A highly conserved region located some sixty residues upstream from the active site glutamate has been selected as a second signature pattern.

Consensus pattern N-x-[LIVMFYWD]-R-[STACN](2)-H-Y-P-x(4)-[LIVMFYWS](2)-x(3)-[DN]-x(2)-G-[LIVMFYW](4) Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern [DENQLF]-[KRVW]-N-[HRY]-[STAPPV]-[SAC]-[LIVMFS](3)-W-[GS]-x(2,3)-N-E [E is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for *Rhizobium meliloti lacZ*.

[1]Henrissat B. *Biochem. J.* 280:309-316(1991).

[2]Schroeder C.J., Robert C., Lenzen G., McKay L.L., Mercenier A. J. *Gen. Microbiol.* 137:369-380(1991).

[3]Gebler J.C., Aebersold R., Withers S.G. *J. Biol. Chem.* 267:11126-11130(1992).

952. (Glyco_hydro_3) Glycosyl hydrolases family 3 active site

PROSITE: PDOC00621. PROSITE cross-reference(s)PS00775; GLYCOSYL_HYDROL_F3

It has been shown [1,2] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:

-Beta glucosidases (EC 3.2.1.21) from the fungi *Aspergillus wentii* (A-3), *Hansenula anomala*, *Kluyveromyces fragilis*, *Saccharomycopsis fibuligera*, (BGL1 and BGL2), *Schizophyllum commune* and *Trichoderma reesei* (BGL1).

-Beta glucosidases from the bacteria *Agrobacterium tumefaciens* (Cbg1), *Butyrivibrio fibrisolvens* (bglA), *Clostridium thermocellum* (bglB), *Escherichia coli* (bglX), *Erwinia chrysanthemi* (bgxA) and *Ruminococcus albus*.

-*Alteromonas* strain O-7 beta-hexosaminidase A (EC 3.2.1.52).

5 -*Bacillus subtilis* hypothetical protein yzbA.

-*Escherichia coli* hypothetical protein ycfO and HI0959, the corresponding *Haemophilus influenzae* protein.

One of the conserved regions in these enzymes is centered on a conserved aspartic acid residue which has been shown [3], in *Aspergillus wentii* beta-glucosidase A3, to be
10 implicated in the catalytic mechanism. This region was used as a signature pattern.

Consensus pattern[LIVM](2)-[KR]-x-[EQK]-x(4)-G-[LIVMFT]-[LIVT]-[LIVMF]-[ST]-D-x(2)-[SGADNI] [D is the active site residue]

Sequences known to belong to this class detected by the patternALL.

15 [1]Henrissat B. Biochem. J. 280:309-316(1991).

[2]Castle L.A., Smith K.D., Morris R.O. J. Bacteriol. 174:1478-1486(1992).

[3]Bause E., Legler G. Biochim. Biophys. Acta 626:459-465(1980).

20 953. GP120 - Envelope glycoprotein GP120

The entry of HIV requires interaction of viral GP120 with Swiss:P01730 and a chemokine receptor on the cell surface. Number of members: 17891

25 [1]Medline: 98303379. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Kwong PD, Wyatt R, Robinson J, Sweet RW, Sodroski J, Hendrickson WA; Nature 1998;393:648-659.

954. (GSP_II_E) Bacterial type II secretion system protein E signature

PROSITE: PDOC00567. PROSITE cross-reference(s) PS00662; T2SP_E

30 A number of bacterial proteins, some of which are involved in a general secretion pathway (GSP) for the export of proteins (also called the type II pathway) [1,2], have been found to be evolutionary related. These proteins are listed below:

-The 'E' protein from the GSP operon of: *Aeromonas* (gene *exeE*); *Erwinia* (gene *outE*); *Escherichia coli* (gene *yheG*); *Klebsiella pneumoniae* (gene *pulE*); *Pseudomonas aeruginosa* (gene *xcpR*); *Vibrio cholerae* (gene *epsE*) and *Xanthomonas campestris* (gene *xpsE*).

-*Agrobacterium tumefaciens* Ti plasmid *virB* operon protein 11. This protein is required for the transfer of T-DNA to plants.

-*Bacillus subtilis* *comG* operon protein 1 which is required for the uptake of DNA by competent *Bacillus subtilis* cells.

-*Aeromonas hydrophila* *tapB*, involved in type IV pilus assembly.

-*Pseudomonas* protein *pilB*, which is essential for the formation of the pili.

-*Pseudomonas aeruginosa* protein twitching mobility protein *pilT*.

-*Neisseria gonorrhoeae* type IV pilus assembly protein *pilF*.

-*Vibrio cholerae* protein *tcpT*, which is involved in the biosynthesis of the *tcp* pilus.

-*Escherichia coli* protein *hofB* (*hopB*).

-*Escherichia coli* hypothetical protein *ygcB*.

-*Escherichia coli* hypothetical protein *yggR*.

These proteins have from 344 (*pilT* and *virB11*) to 568 (*tapB*) amino acids, they are probably cytoplasmically located and, on the basis of the presence of a conserved P-loop region (see <PDOC00017>), probably bind ATP. A region that overlaps the 'B' motif of ATP-binding proteins was selected as a signature pattern.

Consensus pattern[LIVM]-R-x(2)-P-D-x-[LIVM](3)-G-E-[LIVM]-R-D

Sequences known to belong to this class detected by the patternALL, except for *ygcB*.

[1]Salmond G.P.C., Reeves P.J. Trends Biochem. Sci. 18:7-12(1993).

[2]Hobbs M., Mattick J.S. Mol. Microbiol. 10:233-243(1993).

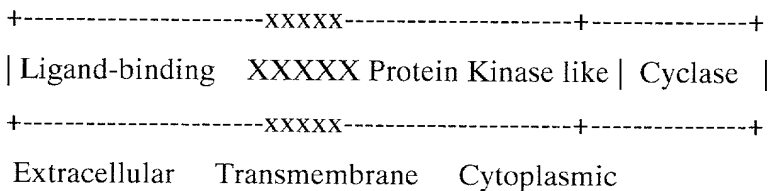
955. (guanylate_cyc) Guanylate cyclases signature

PROSITE: PDOC00425. PROSITE cross-reference(s) PS00452;

GUANYLATE_CYCLASES Guanylate cyclases (EC 4.6.1.2) [1 to 4] catalyze the formation of cyclic GMP (cGMP) from GTP. cGMP acts as an intracellular messenger, activating cGMP dependent kinases and regulating CGMP-sensitive ion channels. The role of cGMP as a second messenger in vascular smooth muscle relaxation and retinal photo-

transduction is well established. Guanylate cyclase is found both in the soluble and particular fraction of eukaryotic cells. The soluble and plasma membrane-bound forms differ in structure, regulation and other properties.

Most currently known plasma membrane-bound forms are receptors for small polypeptides. The topology of such proteins is the following: they have a N-terminal extracellular domain which acts as the ligand binding region, then a transmembrane domain, followed by a large cytoplasmic C-terminal region that can be subdivided into two domains: a protein kinase-like domain that appears important for proper signalling and a cyclase catalytic domain. This topology is schematically represented below.



The known guanylate cyclase receptors are:

- The sea-urchins receptors for speract and resact, which are small peptides that stimulate sperm motility and metabolism.
- The receptors for natriuretic peptides (ANF). Two forms of ANF receptors with guanylate cyclase activity are currently known: GC-A (or ANP-A) which seems specific to atrial natriuretic peptide (ANP), and GC-B (or ANP-B) which seems to be stimulated more effectively by brain natriuretic peptide (BNP) than by ANP.
- The receptor for Escherichia coli heat-stable enterotoxin (GC-C). The endogenous ligand for this intestinal receptor seems to be a small peptide called guanylin.
- Retinal guanylate cyclase (retGC) which probably plays a specific functional role in the rods and/or cones of photoreceptors. It is not known if this protein acts as receptor, but its structure is similar to that of the other plasma membrane-bound GCs.

The soluble forms of guanylate cyclase are cytoplasmic heterodimers. The two subunits, alpha and beta are proteins of from 70 to 82 Kd which are highly related. Two forms of beta subunits are currently known: beta-1 which seems to be expressed in lung and brain, and beta-2 which is more abundant in kidney and liver.

The membrane and cytoplasmic forms of guanylate cyclase share a conserved domain which is probably important for the catalytic activity of the enzyme. Such a domain is also

found twice in the different forms of membrane-bound adenylate cyclases (also known as class-III) [5,6] from mammals, slime mold or *Drosophila*. A consensus pattern was derived from the most conserved region in that domain.

5 Consensus pattern G-V-[LIVM]-x(0,1)-G-x(5)-[FY]-x-[LIVM]-[FYW]-[GS]-[DNTHKW]-
[DNT]-[IV]-[DNTA]-x(5)-[DE]

Sequences known to belong to this class detected by the pattern ALL, except for the sea urchin *Arbacia punctulata* resact receptor which lack this domain.

Note this pattern will detect both domains of adenylate cyclases class-III.

10

[1] Koesling D., Boehme E., Schultz G. *FASEB J.* 5:2785-2791(1991).

[2] Garbers D.L. *New Biol.* 2:499-504(1990).

[3] Garbers D.L. *Cell* 71:1-4(1992).

[4] Yuen P.S.T., Garbers D.L. *Annu. Rev. Neurosci.* 15:193-225(1992).

15 [5] Iyengar R. *FASEB J.* 7:768-775(1993).

[6] Barzu O., Danchin A. *Prog. Nucleic Acid Res. Mol. Biol.* 49:241-283(1994).

956. Hemolysin-type calcium-binding region signature (HemolysinCabinD)

20

Gram-negative bacteria produce a number of proteins which are secreted into the growth medium by a mechanism that does not require a cleaved N-terminal signal sequence. These proteins, while having different functions, seem [1] to share two properties: they bind calcium and they contain a variable number of tandem repeats consisting of a nine amino acid motif rich in glycine, aspartic acid and asparagine. It has been shown [2] that such a domain is involved in the binding of calcium ions in a parallel beta roll structure. The proteins which are currently known to belong to this category are:

25

- Hemolysins from various species of bacteria. Bacterial hemolysins are exotoxins that attack blood cell membranes and cause cell rupture. The hemolysins which are known to contain such a domain are those from: *E. coli* (gene *hlyA*), *A. pleuropneumoniae* (gene *appA*), *A. actinomycetemcomitans* and *P. haemolytica* (leukotoxin) (gene *lktA*).

30

- Cyclolysin from *Bordetella pertussis* (gene *cyaA*). A multifunctional protein which is both an adenylate cyclase and a hemolysin.

780

- Extracellular zinc proteases: serralyisin (EC 3.4.24.40) from *Serratia*, prtB and prtC from *Erwinia chrysanthemi* and aprA from *Pseudomonas aeruginosa*.

- Nodulation protein nodO from *Rhizobium leguminosarum*.

A signature pattern was derived from conserved positions in the sequence of the calcium-binding domain.

Consensus pattern D-x-[LI]-x(4)-G-x-D-x-[LI]-x-G-G-x(3)-D Sequences known to belong to this class detected by the pattern ALL.

Note: This pattern is found once in nodO and the extracellular proteases but up to 5 times in some hemolysin/cyclolysins.

[1] Economou A., Hamilton W.D.O., Johnston A.W.B., Downie J.A. EMBO J. 9:349-354(1990).

[2] Baumann U., Wu S., Flaherty K.M., McKay D.B. EMBO J. 12:3357-3364(1993).

957. Hint module (Hint)

This is an alignment of the Hint module in the Hedgehog proteins. It does not include any Inteins which also possess the Hint module.

Number of members: 36

[1] Hall TM, Porter JA, Young KE, Koonin EV, Beachy PA, Leahy DJ; Medline: 97474313 "Crystal structure of a Hedgehog autoprocessing domain: homology between Hedgehog and self-splicing proteins." Cell 1997;91:85-97.

958. Hydantoinase/oxoprolinase (Hydantoinase)

This family includes the enzymes hydantoinase and oxoprolinase EC:3.5.2.9. Both reactions involve the hydrolysis of 5-membered rings via hydrolysis of their internal imide bonds [1].

Number of members: 14

[1] Ye GJ, Breslow EB, Meister A, Guo-jie GE\$[corrected to Ye GJ]; Medline: 97113037
 "The amino acid sequence of rat kidney 5-oxo-L-prolinase determined by cDNA cloning"
 [published erratum appears in J Biol Chem 1997 Feb 14;272(7):4646] J Biol Chem
 1996;271:32293-32300.

5

959. IMP dehydrogenase / GMP reductase signature (IMPDH_N)

IMP dehydrogenase (EC 1.1.1.205) (IMPDH) catalyzes the rate-limiting reaction of de novo
 GTP biosynthesis, the NAD-dependent reduction of IMP into XMP [1]. Inhibition of IMP
 dehydrogenase activity results in the cessation of DNA synthesis. As IMP dehydrogenase is
 associated with cell proliferation, it is a possible target for cancer chemotherapy. Mammalian
 and bacterial IMPDHs are tetramers of identical chains. There are two IMP dehydrogenase
 isozymes in humans [2].

10

GMP reductase (EC 1.6.6.8) catalyzes the irreversible and NADPH-dependent reductive
 deamination of GMP into IMP [3]. It converts nucleobase, nucleoside and nucleotide
 derivatives of G to A nucleotides, and maintains intracellular balance of A and G nucleotides.

15

IMP dehydrogenase and GMP reductase share many regions of sequence similarity. One of
 these regions is centered on a cysteine residue thought [3] to be involved in binding IMP.
 This region was used as a signature pattern.

20

Consensus pattern[LIVM]-[RK]-[LIVM]-G-[LIVM]-G-x-G-S-[LIVM]-C-x-T [C is the
 putative IMP-binding residue] Sequences known to belong to this class detected by the
 pattern ALL.

25

[1] Collart F.R., Huberman E. J. Biol. Chem. 263:15769-15772(1988).

[2] Natsumeda Y., Ohno S., Kawasaki H., Konno Y., Weber G., Suzuki K. J. Biol. Chem.
 265:5292-5295(1990).

[3] Andrews S.C., Guest J.R. Biochem. J. 255:35-43(1988).

30

960. impB/mucB/samB family (IMS)

782

These proteins are involved in UV protection (Swiss).

Number of members: 38

961. Type II intron maturase (Intron_maturas2)

5

Group II introns use intron-encoded reverse transcriptase, maturase and DNA endonuclease activities for site-specific insertion into DNA [2]. Although this type of intron is self splicing in vitro they require a maturase protein for

10 splicing in vivo. It has been shown that a specific region of the aI2 intron is needed for the maturase function [1]. This region was found to be conserved in group II introns and called domain X [3].

Number of members: 335

[1] Moran JV, Mecklenburg KL, Sass P, Belcher SM, Mahnke D, Lewin A, Perlman P; Medline: 94301788 "Splicing defective mutants of the COXI gene of yeast mitochondrial DNA: initial definition of the maturase domain of the group II intron aI2. Nucleic Acids Res 1994;22:2057-2064.

[2] Guo H, Zimmerly S, Perlman PS, Lambowitz AM; Medline: 98031910 "Group II intron endonucleases use both RNA and protein subunits for recognition of specific sequences in double-stranded DNA." EMBO J 1997;16:6835-6848.

[3] Mohr G, Perlman PS, Lambowitz AM; Medline: 94077696 "Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function." Nucleic Acids Res 1993;21:4991-4997.

25 962. LAGLIDADG endonuclease (Intron_maturase)

[1] Heath PJ, Stephens KM, Monnat RJ Jr, Stoddard BL; Medline: 97331323 "The structure of I-Crel, a group I intron-encoded homing endonuclease." Nat Struct Biol 1997;4:468-476.

[2] Belfort M, Roberts RJ; Medline: 97402526 "Homing endonucleases: keeping the house in order." Nucleic Acids Res 1997;25:3379-3388.

[3] Dalgaard JZ, Klar AJ, Moser MJ, Holley WR, Chatterjee A, Mian IS; Medline: 98026854 "Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases

30

and identification of an intein that encodes a site-specific endonuclease of the HNH family.”
Nucleic Acids Res 1997;25:4626-4638.

Number of members: 220

5

963. Isopentenyl transferase (IPT)

Isopentenyl transferase / dimethylallyl transferase synthesizes isopentenyladenosine 5'-
monophosphate, a cytokinin that induces shoot formation on host plants infected with the Ti
plasmid [1].

10

Number of members: 16

[1] Canaday J, Gerad JC, Crouzet P, Otten L; Medline: 93101133 "Organization and
functional analysis of three T-DNAs from the vitopine Ti plasmid pTiS4." Mol Gen Genet
1992;235:292-303.

15

964. Laminin EGF-like (Domains III and V) (laminin_EGF)

This family is like EGF but has 8 conserved cysteines instead of 6.

Number of members: 501

20

[1] Engel J; Medline: 93041759 "Laminins and other strange proteins." Biochemistry
1992;31:10643-10651.

25 965. Legume lectins signatures (lectin_legA)

Leguminous plants synthesize sugar-binding proteins which are called legume lectins [1,2].
These lectins are generally found in the seeds. The exact function of legume lectins is not
known but they may be involved in the attachment of nitrogen-fixing bacteria to legumes and
in the protection against pathogens. Legume lectins bind calcium and manganese (or other
transition metals).

30

Legume lectins are synthesized as precursor proteins of about 230 to 260 amino acid residues. Some legume lectins are proteolytically processed to produce two chains: beta (which corresponds to the N-terminal) and alpha (C-terminal). The lectin concanavalin A (conA) from jack bean is exceptional in that the two chains are transposed and ligated (by formation of a new peptide bond). The N-terminus of mature conA thus corresponds to that of the alpha chain and the C-terminus to the beta chain.

Two signature patterns were developed specific to legume lectins: the first is located in the C-terminal section of the beta chain and contains a conserved aspartic acid residue important for the binding of calcium and manganese; the second one is located in the N-terminal of the alpha chain.

Consensus pattern [LIV]-[STAG]-V-[DEQV]-[FLI]-D-[ST] [D binds manganese and calcium] Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern [LIV]-x-[EDQ]-[FYWKR]-V-x-[LIVF]-G-[LF]-[ST] Sequences known to belong to this class detected by the pattern ALL.

[1] Sharon N., Lis H. FASEB J. 4:3198-320(1990).

[2] Lis H., Sharon N. Annu. Rev. Biochem. 55:33-37(1986).

966. Malate synthase signature (malate_synthase)

Malate synthase (EC 4.1.3.2) catalyzes the aldol condensation of glyoxylate with acetyl-CoA to form malate - the second step of the glyoxylate bypass, an alternative to the tricarboxylic acid cycle in bacteria, fungi and plants. Malate synthase is a protein of 530 to 570 amino acids whose sequence is highly conserved across species [1]. As a signature pattern, a very conserved region was selected in the central section of the enzyme.

Consensus pattern[KR]-[DENQ]-H-x(2)-G-L-N-x-G-x-W-D-Y-[LIVM]-F Sequences known to belong to this class detected by the pattern ALL.

[1] Bruinenberg P.G., Blaauw M., Kazemier B., Ab G. Yeast 6:245-254(1990).

967. MatK/TrnK amino terminal region (MatK_N)

[1] Mohr G, Perlman PS, Lambowitz AM; Medline: 94077696 "Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function." Nucleic Acids Res 1993;21:4991-4997.

Number of members: 495

968. MOZ/SAS family (MOZ_SAS)

This region of these proteins has been suggested to be homologous to acetyltransferases [1]. However the similarity is not supported by standard sequence analysis.

Number of members: 15

[1] Kamine J, Elangovan B, Subramanian T, Coleman D, Chinnadurai G; Medline: 96182937 "Identification of a cellular protein that specifically interacts with the essential cysteine region of the HIV-1 Tat transactivator." Virology 1996;216:357-366.

[2] Reifsnyder C, Lowell J, Clarke A, Pillus L; Medline: 96376969 "Yeast SAS silencing genes and human genes associated with AML and HIV-1 Tat interactions are homologous with acetyltransferases" [see comments] [published erratum appears in Nat Genet 1997 May;16(1):109] Nat Genet 1996;14:42-49.

969. mRNA capping enzyme (mRNA_cap_enzyme)

[1] Hakansson K, Doherty AJ, Shuman S, Wigley DB; Medline: 97304383 "X-ray crystallography reveals a large conformational change during guanyl transfer by mRNA capping enzymes." Cell 1997;89:545-553.

Number of members: 7

970. DNA mismatch repair proteins mutS family signature (MutS_C)

Mismatch repair contributes to the overall fidelity of DNA replication [1]. It involves the correction of mismatched base pairs that have been missed by the proofreading element of the DNA polymerase complex. The sequence of some proteins involved in mismatch repair in different organisms have been found to be evolutionary related [2,3]. One of these families is called mutS [4,E1], it consists of:

- Prokaryotic protein mutS protein (also called hexA in *Streptococcus pneumoniae*). Muts is thought to carry out the mismatch recognition step of DNA repair.
- Eukaryotic MSH1, which is involved in mitochondrial DNA repair.
- Eukaryotic MSH2, which is involved in nuclear postreplication mismatch repair. MSH2 heterodimerizes with MSH6. In man, MSH2 is involved in a form of familial hereditary nonpolyposis colon cancer (HNPCC).
- Eukaryotic MSH3, which is probably involved in the repair of large loops.
- Eukaryotic MSH4, which is involved in meiotic recombination.
- Eukaryotic MSH5, which is involved in meiotic recombination.
- Eukaryotic MSH6 (also known as G/T mismatch binding protein), a DNA-repair protein that binds to G/T mismatches through heterodimerization with MSH2.
- Prokaryotic protein mutS2 whose function is not yet known.
- A coral (*Sarcophyton glaucum*) mitochondrial encoded mutS-like protein.

As a signature pattern for this class of mismatch repair proteins a region rich in glycine and negatively charged residues was selected. This region is found in the C-terminal section of these proteins; about 80 residues to the C-terminal of an ATP-binding site motif 'A' (P-loop) (see <PDOC00017>).

Consensus pattern[ST]-[LIVMF]-x-[LIVM]-x-D-E-[LIVMFY]-[GC]-[RKH]-G-[GST]- x(4)-
G Sequences known to belong to this class detected by the pattern ALL, except for mutS2.

[1] Modrich P. Annu. Rev. Biochem. 56:435-466(1987).

[2] Haber L.T., Walker G.C. EMBO J. 10:2707-2715(1991).

[3] New L., Liu K., Crouse G.F. Mol. Gen. Genet. 239:97-108(1993).

[4] Eisen J.A. Nucleic Acids Res. 26:4291-4300(1998).

971. MutS family, N-terminal putative DNA binding domain (MutS_N)

This family consists of the N-terminal region of proteins in the mutS family of DNA mismatch repair proteins and is found associated with MutS_C located in the C-terminal region. The mutS family of proteins is named after the salmonella typhimurium MutS protein involved in mismatch repair; other members of the family included the eukaryotic MSH 1,2,3,4,5 and 6 proteins. These have various roles in DNA repair and recombination. Human MSH has been implicated in non-polyposis colorectal carcinoma (HNPCC) and is a mismatch binding protein [2]. The aligned region corresponds in part with domains A1, A2 (which may bind DNA) and B (which binds dsDNA in vitro) from T. thermophilus MutS as characterised in [1].

10 Number of members: 43

972. Domain in Myosin and Kinesin Tails (MyTH4)

Domain present twice in myosin-VIIa, and also present in 3 other myosins.

[1] Chen ZY, Hasson T, Kelley PM, Schwender BJ, Schwartz MF, Ramakrishnan M, Kimberling WJ, Mooseker MS, Corey DP; Medline: 97038686 "Molecular cloning and domain structure of human myosin-VIIa, the gene product defective in Usher syndrome 1B." Genomics 1996;36:440-448.

20 Number of members: 21

973. Sodium and potassium ATPases beta subunits signatures (Na_K-ATPase)

25 The sodium pump (Na⁺,K⁺ ATPase), located in the plasma membrane of all animal cells [1], is an heterotrimer of a catalytic subunit (alpha chain), a glycoprotein subunit of about 34 Kd (beta chain) and a small hydrophobic protein of about 6 Kd. The beta subunit seems [2] to regulate, through the assembly of alpha/beta heterodimers, the number of sodium pumps transported to the plasma membrane.

30

Structurally the beta subunit is composed of a charged cytoplasmic domain of about 35 residues, followed by a transmembrane region, and a large extracellular domain that contains

788

three disulfide bonds and glycosylation sites. This structure is schematically represented in the figure below.

```
+-----+ +---+ +-----+ |||||
XXXXXXXXXXXXXXXXXXXXCXXXXCXCXXCXXXXXXCXXXXXXXXXXCXXXX
5  ****  <-Cyt-><TM><-----Extracellular----->
```

'C': conserved cysteine involved in a disulfide bond.

'*': position of the patterns.

- 10 Two isoforms of the beta subunit (beta-1 and beta-2) are currently known; they share about 50% sequence identity. Gastric (K⁺, H⁺) ATPase (proton pump) responsible for acid production in the stomach consist of two subunits [3]; the beta chain is highly similar to the sodium pump beta subunits. Two signature patterns were developed for beta subunits. The first is located in the cytoplasmic domain, while the second is found in the extracellular domain and contains two of the cysteines involved in disulfide bonds.

Consensus pattern [FYW]-x(2)-[FYW]-x-[FYW]-[DN]-x(6)-[LIVM]-G-R-T-x(3)-W
Sequences known to belong to this class detected by the pattern ALL.

- 20 Consensus pattern [RK]-x(2)-C-[RKQWI]-x(5)-L-x(2)-C-[SA]-G [The two C's are involved in disulfide bonds] Sequences known to belong to this class detected by the pattern ALL, except for the beta subunit of the sodium pump of brine shrimp whose sequence is highly divergent in that region.

- 25 [1] Horisberger J.D., Lemas V., Krahenbul J.P., Rossier B.C. Annu. Rev. Physiol. 53:565-584(1991).
[2] McDonough A.A., Gerring K., Farley R.A. FASEB J. 4:1598-1605(1990).
[3] Toh B.-H., Gleeson P.A., Simpson R.J., Moritz R.L., Callaghan J.M., Goldkorn I., Jones C.M., Martinelli T.M., Mu F.-T., Humphris D.C., Pettitt J.M., Mori Y., Masuda T.,
30 Sobieszczuk P., Weinstock J., Mantamadiotis T., Baldwin G.S. Proc. Natl. Acad. Sci. U.S.A. 87:6418-6422(1990).

974. Respiratory-chain NADH dehydrogenase subunit 1 signatures (NADHdh)

Respiratory-chain NADH dehydrogenase (EC 1.6.5.3) [1,2] (also known as complex I or NADH-ubiquinone oxidoreductase) is an oligomeric enzymatic complex located in the inner mitochondrial membrane which also seems to exist in the chloroplast and in cyanobacteria (as a NADH-plastoquinone oxidoreductase). Among the 25 to 30 polypeptide subunits of this bioenergetic enzyme complex there are fifteen which are located in the membrane part, seven of which are encoded by the mitochondrial and chloroplast genomes of most species. The most conserved of these organelle-encoded subunits is known as subunit 1 (gene ND1 in mitochondrion, and NDH1 in chloroplast) and seems to contain the ubiquinone binding site.

The ND1 subunit is highly similar to subunit 4 of *Escherichia coli* formate hydrogenlyase (gene hycD), subunit C of hydrogenase-4 (gene hyfC). *Paracoccus denitrificans* NQO8 and *Escherichia coli* nuoH NADH-ubiquinone oxidoreductase subunits also belong to this family [3]. Two signature patterns were developed based on conserved regions of this subunit.

Consensus pattern G-[LIVMFYKRS]-[LIVMAGP]-Q-x-[LIVMFY]-x-D-[AGIM]-[LIVMFTA]-K-[LVMYST]-[LIVMFYG]-x-[KR]-[EQG] Sequences known to belong to this class detected by the pattern ALL, except for watermelon and *Leishmania* ND1.

Consensus pattern P-F-D-[LIVMFYQ]-[STAGPVM]-E-[GAC]-E-x-[EQ]-[LIVMS]-x(2)-G Sequences known to belong to this class detected by the pattern ALL, except for *Chlamydomonas reinhardtii* and *Pisaster ochraceus* ND1, and tobacco NDH1.

[1] Ragan C.I. Curr. Top. Bioenerg. 15:1-36(1987).

[2] Weiss H., Friedrich T., Hofhaus G., Preis D. Eur. J. Biochem. 197:563-576(1991).

[3] Weidner U., Geier S., Ptock A., Friedrich T., Leif H., Weiss H. J. Mol. Biol. 233:109-122(1993).

975. Nickel-dependent hydrogenases large subunit signatures (NiFeSe_Hases)

Hydrogenases are enzymes that catalyze the reversible activation of hydrogen and which occur widely in prokaryotes as well as in some eukaryotes. There are various types of hydrogenases, but all of them seem to contain at least one iron-sulfur cluster. They can be

broadly divided into two groups: hydrogenases containing nickel and, in some cases, also selenium (the [NiFe] and [NiFeSe] hydrogenases) and those lacking nickel (the [Fe] hydrogenases).

5 The [NiFe] and [NiFeSe] hydrogenases are heterodimer that consist of a small subunit that contains a signal peptide and a large subunit. All the known large subunits seem to be evolutionary related [1]; they contain two Cys-x-x- Cys motifs; one at their N-terminal end; the other at their C-terminal end. These four cysteines are involved in the binding of nickel [2]. In the [NiFeSe] hydrogenases the first cysteine of the C-terminal motif is a
10 selenocysteine which has experimentally been shown to be a nickel ligand [3]. Two patterns were developed which are centered on the Cys-x-x-Cys motifs.

Alcaligenes eutrophus possess a NAD-reducing cytoplasmic hydrogenase (hoxS) [4]; this enzyme is composed of four subunits. Two of these subunits (beta and delta) are responsible
15 for the hydrogenase reaction and are evolutionary related to the large and small subunits of membrane-bound hydrogenases. The alpha subunit of coenzyme F420 hydrogenase (EC 1.12.99.1) (FRH) from archaeobacterial methanogens also belongs to this family.

Consensus pattern R-G-[LIVMF]-E-x(15)-[QESM]-R-x-C-G-[LIVM]-C [The two C's are
20 nickel ligands] Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern [FY]-D-P-C-[LIM]-[ASG]-C-x(2,3)-H [The two C's are nickel ligands]
Sequences known to belong to this class detected by the pattern ALL.

25 [1] Menon N.K., Robbins J., Peck H.D. Jr., Chatelus C.Y., Choi E.-S., Przybyla A.E. J. Bacteriol. 172:1969-1977(1990).

[2] Volbeda A., Charon M.-H., Piras C., Hatchikian E.C., Frey M., Fontecilla-Camps J.C. Nature 373:580-587(1995).

[3] Eidsness M.K., Scott R.A., Prickrill B., der Vartanian D.V., LeGall J., Moura I., Moura J.J.G., Peck H.D. Jr. Proc. Natl. Acad. Sci. U.S.A. 86:147-151(1989).
30

[4] Tran-Betcke A., Warnecke U., Boecker C., Zaborosch C., Friedrich B. J. Bacteriol. 172:2920-2929(1990).

791

976. NADH-Ubiquinone oxidoreductase (complex I), chain 5 C-terminus (oxidored_q1_C)

This sub-family represents a carboxyl terminal extension of oxidored_q1. Only NADH-Ubiquinone chain 5 from chloroplasts are in this family. This sub-family is part of complex I which catalyses the transfer of two electrons from NADH to ubiquinone in a reaction that is associated with proton translocation across the membrane.

Number of members: 572

[1] Walker JE; Medline: 93110040 "The NADH:ubiquinone oxidoreductase (complex I) of respiratory chains." Q Rev Biophys 1992;25:253-324.

977. NADH-Ubiquinone oxidoreductase (complex I), chain 5 N-terminus (oxidored_q1_N)

This sub-family represents an amino terminal extension of oxidored_q1. Only NADH-Ubiquinone chain 5 and eubacterial chain L are in this family. This sub-family is part of complex I which catalyses the transfer of two electrons from NADH to ubiquinone in a reaction that is associated with proton translocation across the membrane.

Number of members: 546

[1] Walker JE; Medline: 93110040 "The NADH:ubiquinone oxidoreductase (complex I) of respiratory chains." Q Rev Biophys 1992;25:253-324.

978. oxidored_q2. NADH-UBIQUINONE OXIDOREDUCTASE CHAIN 4L (EC 1.6.5.3). ND4L OR NAD4L. Arabidopsis thaliana (Mouse-ear cress). Mitochondrion. OC Eukaryota; Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; Rosidae; eurosids II; Brassicales; Brassicaceae; Arabidopsis. CATALYTIC ACTIVITY: NADH + UBIQUINONE = NAD(+) + UBIQUINOL.

[1] SEQUENCE FROM N.A. MEDLINE; 93156682. Brandt P., Sunkel S., Unseld M., Brennicke A., Knoop V.; "The nad4L gene is encoded between exon c of nad5 and orf25 in the Arabidopsis mitochondrial genome."; Mol. Gen. Genet. 236:33-38(1992).

[2] SEQUENCE FROM N.A. STRAIN=CV. COLUMBIA; MEDLINE; 97141919 Unseld M., Marienfeld J.R., Brandt P., Brennicke A.; "The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides."; Nat. Genet. 15:57-61(1997).

5 979. oxidored_q4. Protein name NADH-PLASTOQUINONE OXIDOREDUCTASE CHAIN 3, CHLOROPLAST. Synonym(s)EC 1.6.5.3. Gene name(s)NDHC OR NDH3 From *Zea mays* (Maize) Encoded on Chloroplast. Taxonomy Eukaryota; Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; *Zea*.

CATALYTIC ACTIVITY: NADH + PLASTOQUINONE = NAD(+) +

10 PLASTOQUINOL.

SIMILARITY: BELONGS TO THE COMPLEX I SUBUNIT 3 FAMILY.

[1] SEQUENCE FROM N.A. MEDLINE; 89281491. Steinmueller K., Ley A.C., Steinmetz A.A., Sayre R.T., Bogorad L.; "Characterization of the *ndhC-psbG-ORF157/159* operon of maize plastid DNA and of the cyanobacterium *Synechocystis* sp. PCC6803."; Mol. Gen. Genet. 216:60-69(1989).

[2] SEQUENCE FROM N.A. MEDLINE; 95395841. Maier R.M., Neckermann K., Igloi G.L., Koessel H.; "Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing."; J. Mol. Biol. 251:614-628(1995).

980. PAC: PAC motif

PAC motif occurs C-terminal to a subset of all known PAS motifs. It is proposed to contribute to the PAS domain fold [3]. Number of members: 181

[1] Medline: 97446881 PAS domain S-boxes in archaea, bacteria and sensors for oxygen and redox. Zhulin IB, Taylor BL, Dixon R; Trends Biochem Sci 1997;22:331-333.

[2] Medline: 95275818. 1.4 A structure of photoactive yellow protein, a cytosolic photoreceptor: unusual fold, active site, and chromophore. Borgstahl GE, Williams DR, Getzoff ED; Biochemistry 1995;34:6278-6287.

[3] Medline: 98044337. PAS: a multifunctional domain family comes to light. Ponting CP, Aravind L; Curr Biol 1997;7:674-677.

793

981. PARP: Poly(ADP-ribose) polymerase catalytic region.

Poly(ADP-ribose) polymerase catalyses the covalent attachment of ADP-ribose units from NAD⁺ to itself and to a limited number of other DNA binding proteins, which decreases their affinity for DNA. Poly(ADP-ribose) polymerase is a regulatory component induced by DNA damage.

The carboxyl-terminal region is the most highly conserved region of the protein. Experiments have shown that a carboxyl 40 kDa fragment is still catalytically active [2]. Number of members: 19

[1] Medline: 96353841 Structure of the catalytic fragment of poly(AD-ribose) polymerase from chicken. Ruf A, Mennissier de Murcia J, de Murcia G, Schulz GE; Proc Natl Acad Sci U S A 1996;93:7481-7485.

[2] Medline: 93293867 The carboxyl-terminal domain of human poly(ADP-ribose) polymerase. Overproduction in Escherichia coli, large scale purification, and characterization. Simonin F, Hofferer L, Panzeter PL, Muller S, de Murcia G, Althaus FR; J Biol Chem 1993;268:13454-13461.

982. PC_rep: Proteasome/cyclosome repeat

[1] Medline: 97348748 A repetitive sequence in subunits of the 26S proteasome and 20S cyclosome (anaphase-promoting complex). Lupas A, Baumeister W, Hofmann K; Trends Biochem Sci 1997;22:195-196.

Number of members: 112

983. Peptidase_M1: Peptidase family M1

Members of this family are aminopeptidases. The members differ widely in specificity, hydrolysing acidic, basic or neutral N-terminal residues. This family includes leukotriene-A4 hydrolase Swiss:P09960, this enzyme also has an aminopeptidase activity [1]. Number of members: 72

[1] Medline: 95405261 Evolutionary families of metallopeptidases. Rawlings ND, Barrett AJ; Meth Enzymol 1995;248:183-228.

984. Neutral zinc metallopeptidases, zinc-binding region signature (Peptidase_M8)

PROSITE cross-reference(s) PS00142; ZINC_PROTEASE

The majority of zinc-dependent metallopeptidases (with the notable exception of the carboxypeptidases) share a common pattern of primary structure [1,2,3] in the part of their sequence involved in the binding of zinc, and can be grouped together as a superfamily, known as the metzincins, on the basis of this sequence similarity. They can be classified into a number of distinct families [4,E1] which are listed below along with the proteases which are currently known to belong to these families.

Family M1

- Bacterial aminopeptidase N (EC 3.4.11.2) (gene pepN).
- Mammalian aminopeptidase N (EC 3.4.11.2).
- Mammalian glutamyl aminopeptidase (EC 3.4.11.7) (aminopeptidase A). It may play a role in regulating growth and differentiation of early B-lineage cells.
- Yeast aminopeptidase yscII (gene APE2).
- Yeast alanine/arginine aminopeptidase (gene AAP1).
- Yeast hypothetical protein YIL137c.
- Leukotriene A-4 hydrolase (EC 3.3.2.6). This enzyme is responsible for the hydrolysis of an epoxide moiety of LTA-4 to form LTB-4; it has been shown that it binds zinc and is capable of peptidase activity.

Family M2

- Angiotensin-converting enzyme (EC 3.4.15.1) (dipeptidyl carboxypeptidase I) (ACE) the enzyme responsible for hydrolyzing angiotensin I to angiotensin II. There are two forms of ACE: a testis-specific isozyme and a somatic isozyme which has two active centers.

Family M3

- Thimet oligopeptidase (EC 3.4.24.15), a mammalian enzyme involved in the cytoplasmic degradation of small peptides.
- Neurolysin (EC 3.4.24.16) (also known as mitochondrial oligopeptidase M or microsomal endopeptidase).
- Mitochondrial intermediate peptidase precursor (EC 3.4.24.59) (MIP). It is involved the second stage of processing of some proteins imported in the mitochondrion.
- Yeast saccharolysin (EC 3.4.24.37) (proteinase yscD).

795

- *Escherichia coli* and related bacteria dipeptidyl carboxypeptidase (EC 3.4.15.5) (gene dcp).
- *Escherichia coli* and related bacteria oligopeptidase A (EC 3.4.24.70) (gene opdA or prlC).
- Yeast hypothetical protein YKL134c.

5 Family M4

- Thermostable thermolysins (EC 3.4.24.27), and related thermolabile neutral proteases (bacillolysins) (EC 3.4.24.28) from various species of *Bacillus*.
- Pseudolysin (EC 3.4.24.26) from *Pseudomonas aeruginosa* (gene lasB).
- Extracellular elastase from *Staphylococcus epidermidis*.

10

- Extracellular protease prt1 from *Erwinia carotovora*.
- Extracellular minor protease smp from *Serratia marcescens*.
- Vibriolysin (EC 3.4.24.25) from various species of *Vibrio*.
- Protease prtA from *Listeria monocytogenes*.
- Extracellular proteinase proA from *Legionella pneumophila*.

15

Family M5

- Mycolysin (EC 3.4.24.31) from *Streptomyces cacaoi*.

Family M6

20

- Immune inhibitor A from *Bacillus thuringiensis* (gene ina). Ina degrades two classes of insect antibacterial proteins, attacins and cecropins.

Family M7

- *Streptomyces* extracellular small neutral proteases

25

Family M8

- Leishmanolysin (EC 3.4.24.36) (surface glycoprotein gp63), a cell surface protease from various species of *Leishmania*.

30

Family M9

- Microbial collagenase (EC 3.4.24.3) from *Clostridium perfringens* and *Vibrio alginolyticus*.

Family M10A

- Serralysin (EC 3.4.24.40), an extracellular metalloprotease from *Serratia*.
- Alkaline metalloproteinase from *Pseudomonas aeruginosa* (gene *aprA*).
- Secreted proteases A, B, C and G from *Erwinia chrysanthemi*.

5 - Yeast hypothetical protein YIL108w.

Family M10B

- Mammalian extracellular matrix metalloproteinases (known as matrixins) [5]: MMP-1 (EC 3.4.24.7) (interstitial collagenase), MMP-2 (EC 3.4.24.24) (72 Kd gelatinase), MMP-9 (EC 3.4.24.35) (92 Kd gelatinase), MMP-7 (EC 3.4.24.23) (matrylisin), MMP-8 (EC 3.4.24.34) (neutrophil collagenase), MMP-3 (EC 3.4.24.17) (stromelysin-1), MMP-10 (EC 3.4.24.22) (stromelysin-2), and MMP-11 (stromelysin-3), MMP-12 (EC 3.4.24.65) (macrophage metalloelastase).
- Sea urchin hatching enzyme (envelysin) (EC 3.4.24.12). A protease that allows the embryo to digest the protective envelope derived from the egg extracellular matrix.
- Soybean metalloendoproteinase 1.

Family M11

- *Chlamydomonas reinhardtii* gamete lytic enzyme (GLE).

Family M12A

- Astacin (EC 3.4.24.21), a crayfish endoprotease.
- Meprin A (EC 3.4.24.18), a mammalian kidney and intestinal brush border metalloendopeptidase.
- Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation and which expresses metalloendopeptidase activity. The *Drosophila* homolog of BMP-1 is the dorsal-ventral patterning protein *tolloid*.
- Blastula protease 10 (BP10) from *Paracentrotus lividus* and the related protein SpAN from *Strongylocentrotus purpuratus*.
- *Caenorhabditis elegans* protein *toh-2*.
- *Caenorhabditis elegans* hypothetical protein F42A10.8.
- Choriolytins L and H (EC 3.4.24.67) (also known as embryonic hatching proteins LCE and HCE) from the fish *Oryzias latipes*. These proteases participate in the breakdown

of the egg envelope, which is derived from the egg extracellular matrix, at the time of hatching.

Family M12B

- 5 - Snake venom metalloproteinases [6]. This subfamily mostly groups proteases that act in hemorrhage. Examples are: adamalysin II (EC 3.4.24.46), atrolysin C/D (EC 3.4.24.42), atrolysin E (EC 3.4.24.44), fibrolase (EC 3.4.24.72), trimereylisin I (EC 3.4.25.52) and II (EC 3.4.25.53).
- 10 - Mouse cell surface antigen MS2.

Family M13

- 15 - Mammalian neprilysin (EC 3.4.24.11) (neutral endopeptidase) (NEP).
- 20 - Endothelin-converting enzyme 1 (EC 3.4.24.71) (ECE-1), which process the precursor of endothelin to release the active peptide.
- 25 - Kell blood group glycoprotein, a major antigenic protein of erythrocytes. The Kell protein is very probably a zinc endopeptidase.
- 30 - Peptidase O from *Lactococcus lactis* (gene pepO).

Family M27

- 35 - Clostridial neurotoxins, including tetanus toxin (TeTx) and the various botulinum toxins (BoNT). These toxins are zinc proteases that block neurotransmitter release by proteolytic cleavage of synaptic proteins such as synaptobrevins, syntaxin and SNAP-25 [7,8].

Family M30

- 40 - *Staphylococcus hyicus* neutral metalloprotease.

Family M32

- 45 - Thermostable carboxypeptidase 1 (EC 3.4.17.19) (carboxypeptidase Taq), an enzyme from *Thermus aquaticus* which is most active at high temperature.

Family M34

- Lethal factor (LF) from *Bacillus anthracis*, one of the three proteins composing the anthrax toxin.

Family M35

- 5 - Deuterolysin (EC 3.4.24.39) from *Penicillium citrinum* and related proteases from various species of *Aspergillus*.

Family M36

- Extracellular elastinolytic metalloproteinases from *Aspergillus*.

10

From the tertiary structure of thermolysin, the position of the residues acting as zinc ligands and those involved in the catalytic activity are known. Two of the zinc ligands are histidines which are very close together in the sequence; C-terminal to the first histidine is a glutamic acid residue which acts as a nucleophile and promotes the attack of a water molecule on the carbonyl carbon of the substrate. A signature pattern which includes the two histidine and the glutamic acid residues is sufficient to detect this superfamily of proteins.

15

Consensus pattern[GSTALIVN]-x(2)-H-E-[LIVMFYW]-{DEHRKP}-H-x-
[LIVMFYWGSPQ]

20

[The two H's are zinc ligands] [E is the active site residue]

Sequences known to belong to this class detected by the patternALL, except for members of families M5, M7 and M11.

Other sequence(s) detected in SWISS-PROT57; including *Neurospora crassa* conidiation-specific protein 13 which could be a zinc-protease.

25

[1]Jongeneel C.V., Bouvier J., Bairoch A. FEBS Lett. 242:211-214(1989).

[2]Murphy G.J.P., Murphy G., Reynolds J.J. FEBS Lett. 289:4-7(1991).

[3]Bode W., Grams F., Reinemer P., Gomis-Rueth F.-X., Baumann U., McKay D.B., Stoecker W. Zoology 99:237-246(1996).

30

[4]Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).

[5]Woessner J. Jr. FASEB J. 5:2145-2154(1991).

[6]Hite L.A., Fox J.W., Bjarnason J.B. Biol. Chem. Hoppe-Seyler 373:381-385(1992).

[7]Montecucco C., Schiavo G. Trends Biochem. Sci. 18:324-327(1993).

[8]Niemann H., Blasi J., Jahn R. Trends Cell Biol. 4:179-185(1994).

985. PHO4: Phosphate transporter family

This family includes PHO-4 from *Neurospora crassa* which is a Na(+)-phosphate symporter [1]. This family also contains the leukemia virus receptor Swiss:Q08344. Number of members: 41

[1] Medline: 95249577 Repressible cation-phosphate symporters in *Neurospora crassa*. Versaw WK, Metzenberg RL; Proc Natl Acad Sci U S A 1995;92:3884-3887.

986. Photosynthetic reaction center proteins signature (photoRC)

PROSITE cross-reference(s): PS00244; REACTION_CENTER

In the photosynthetic reaction center of purple bacteria, two homologous integral membrane proteins, L(ight) and M(edium), are known to be essential to the light-mediated water-splitting process. In the photosystem II of eukaryotic chloroplasts two related proteins are involved: the D1 (psbA) and D2 proteins (psbD). These four types of protein probably evolved from a common ancestor [see 1,2 for recent reviews].

A signature pattern was developed which include two conserved histidine residues. In L and M chains, the first histidine is a ligand of the magnesium ion of the special pair bacteriochlorophyll, the second is a ligand of a ferrous non-heme iron atom. In photosystem II these two histidines are thought to play a similar role.

Consensus pattern[NQH]-x(4)-P-x-H-x(2)-[SAG]-x(11)-[SAGC]-x-H-[SAG](2)

[The first H is a magnesium ligand] [The second H is a iron ligand]

Sequences known to belong to this class detected by the patternALL, except for broad bean psbA which has Gln instead of the second His.

[1]Michel H., Deisenhofer J. Biochemistry 27:1-7(1988).

[2]Barber J. Trends Biochem. Sci. 12:321-326(1987).

987. phytochrome: Phytochrome region

800

This family contains a region specific to phytochrome proteins. Number of members:
145

988. PI3K_C2: C2 domain

- 5 Phosphoinositide 3-kinase region postulated to contain a C2 domain. Outlier of C2 family.
Number of members: 39

[1] Medline: 97388296 Using structure to define the function of phosphoinositide 3-kinase family members. Domin J, Waterfield MD; FEBS Lett 1997;410:91-95.

- 10 [2] Medline: 97398940 Phosphoinositide 3-kinases: a conserved family of signal transducers. Vanhaesebroeck B, Leever SJ, Panayotou G, Waterfield MD; Trends Biochem Sci 1997;22:267-272.

989. PI3Ka: Phosphoinositide 3-kinase family, accessory domain (PIK domain)

- 15 PIK domain is conserved in all PI3 and PI4-kinases. Its role is unclear but it has been suggested [2] to be involved in substrate presentation.

Number of members: 47

[1] Medline: 97388296 Using structure to define the function of phosphoinositide 3-kinase family members. Domin J, Waterfield MD; FEBS Lett 1997;410:91-95.

- 20 [2] Medline: 94069320 Phosphatidylinositol 4-kinase: gene structure and requirement for yeast cell viability. Flanagan CA, Schnieders EA, Emerick AW, Kunisawa R, Admon A, Thorner J; Science 1993;262:1444-1448.

- 25 990. P-II protein signatures

PROSITE cross-reference(s): PS00496; PII_GLNB_UMP, PS00638; PII_GLNB_CTER

The P-II protein (gene glnB) is a bacterial protein important for the control of glutamine synthetase [1,2,3]. In nitrogen-limiting conditions, when the ratio of glutamine to 2-ketoglutarate decreases, P-II is uridylylated on a tyrosine residue to form P-II-UMP. P-II-UMP allows the deadenylation of glutamine synthetase (GS), thus activating the enzyme. Conversely, in nitrogen excess, P-II-UMP is deuridylylated and then promotes the adenylation of GS. P-II also indirectly controls the transcription of the GS gene (glnA) by preventing NR-

30

801

II (ntrB) to phosphorylate NR-I (ntrC) which is the transcriptional activator of glnA. Once P-II is uridylylated, these events are reversed.

P-II is a protein of about 110 amino acid residues extremely well conserved. The tyrosine which is uridylylated is located in the central part of the protein.

In cyanobacteria, P-II seems to be phosphorylated on a serine residue rather than being uridylylated.

In methanogenic archaeobacteria, the nitrogenase iron protein gene (nifH) is followed by two open reading frames highly similar to the eubacterial P-II protein [4]. These proteins could be involved in the regulation of nitrogen fixation.

In the red alga, *Porphyra purpurea*, there is a glnB homolog encoded in the chloroplast genome.

Other proteins highly similar to glnB are:

- *Bacillus subtilis* protein nrgB [5].
- *Escherichia coli* hypothetical protein ybaI [6].

Two signature patterns were developed for P-II protein. The first one is a conserved stretch (in eubacteria) of six residues which contains the uridylylated tyrosine, the other is derived from a conserved region in the C-terminal part of the P-II protein.

Consensus pattern Y-[KR]-G-[AS]-[AE]-Y [The second Y is uridylylated]

Sequences known to belong to this class detected by the pattern ALL glnB's from eubacteria.

Consensus pattern [ST]-x(3)-G-[DY]-G-[KR]-[IV]-[FW]-[LIVM]-x(2)-[LIVM]

[1] Magasanik B. *Biochimie* 71:1005-1012(1989).

[2] Holtel A., Merrick M. *Mol. Gen. Genet.* 215:134-138(1988).

[3]Cheah E., Carr P.D., Suffolk P.M., Vasuvedan S.G., Dixon N.E., Ollis D.L. Structure 2:981-990(1994).

[4]Sibold L., Henriquet M., Possot O., Aubert J.-P. Res. Microbiol. 142:5-12(1991).

[5]Wray L.V. Jr., Atkinson M.R., Fisher S.H. J. Bacteriol. 176:108-114(1994).

5 [6]Allikmets R., Gerrard B.C., Court D., Dean M.C. Gene 136:231-236(1993).

991. PIP5K: Phosphatidylinositol-4-phosphate 5-Kinase

This family contains a region from the common kinase core found in the type I phosphatidylinositol-4-phosphate 5-kinase (PIP5K) family as described in [1]. The family
10 consists of various type I, II and III PIP5K enzymes. PIP5K catalyses the formation of phosphoinositol-4,5-bisphosphate via the phosphorylation of phosphatidylinositol-4-phosphate a precursor in the phosphoinositide signaling pathway. Number of members: 33

[1] Medline: 98204859. Type I phosphatidylinositol-4-phosphate 5-kinases. Cloning of the third isoform and deletion/substitution analysis of members of this novel lipid kinase family. Ishihara H, Shibasaki Y, Kizuki N, Wada T, Yazaki Y, Asano T, Oka Y; J Biol Chem 1998;273:8741-8748.

[2] Medline: 97115834 Type I phosphatidylinositol-4-phosphate 5-kinases are distinct members of this novel lipid kinase family. Loijens JC, Anderson RA; J Biol Chem 1996 20;271:32937-32943.

992. PolyA_pol: Poly A polymerase family

This family includes nucleic acid independent RNA polymerases, such as Poly(A) polymerase, which adds the poly (A) tail to mRNA EC:2.7.7.19. This family also includes the
25 tRNA nucleotidyltransferase that adds the CCA to the 3' of the tRNA EC:2.7.7.25. Number of members: 31

[1] Medline: 93066242 Identification of the gene for an Escherichia coli poly(A) polymerase. Cao GJ, Sarkar N; Proc Natl Acad Sci U S A 1992;89:10380-10384.

993. Photosystem I psaA and psaB proteins signature (psaA_psaB)
PROSITE cross-reference(s)PS00419; PHOTOSYSTEM_I_PSAAB

Photosystem I (PSI) [1] is an integral membrane protein complex that uses light energy to mediate electron transfer from plastocyanin to ferredoxin. PSI is found in the chloroplast of plants and cyanobacteria. The electron transfer components of the reaction center of PSI are a primary electron donor P-700 (chlorophyll dimer) and five electron acceptors: A0 (chlorophyll), A1 (a phylloquinone) and three 4Fe-4S iron-sulfur centers: Fx, Fa, and Fb.

PsaA and psaB, two closely related proteins, are involved in the binding of P700, A0, A1, and Fx. psaA and psaB are both integral membrane proteins of 730 to 750 amino acids that seem to contain 11 transmembrane segments. The Fx 4Fe-4S iron-sulfur center is bound by four cysteines; two of these cysteines are provided by the psaA protein and the two others by psaB. The two cysteines in both proteins are proximal and located in a loop between the ninth and tenth transmembrane segments. A leucine zipper motif seems to be present [2] downstream of the cysteines and could contribute to dimerization of psaA/psaB.

The signature pattern for these proteins is based on the perfectly conserved region that includes the two iron-sulfur binding cysteines.

Consensus pattern C-D-G-P-G-R-G-G-T-C [The two C's bind the iron-sulfur center]

[1] Golbeck J.H. Biochim. Biophys. Acta 895:167-204(1987).

[2] Webber A.N., Malkin R. FEBS Lett. 264:1-14(1990).

994. PSBH: Photosystem II 10 kDa phosphoprotein

This protein is phosphorylated in a light dependent reaction.

Number of members: 20

995. PsbJ

This family consists of the photosystem II reaction center protein PsbJ from plants and Cyanobacteria. In *Synechocystis* sp. PCC 6803 PsbJ regulates the number of photosystem II centers in thylakoid membranes, it is a predicted 4kDa protein with one membrane spanning domain [1]. Number of members: 20

[1] Medline: 93131892. Genetic and immunological analyses of the cyanobacterium

Synechocystis sp. PCC 6803 show that the protein encoded by the psbJ gene regulates the

number of photosystem II centers in thylakoid membranes. Lind LK, Shukla VK, Nyhus KJ, Pakrasi HB; *J Biol Chem* 1993;268:1575-1579.

996. PSBT: Photosystem II reaction centre T protein

5 The exact function of this protein is unknown. It probably consists of a single transmembrane spanning helix. The Swiss:P37256 protein, appears to be (i) a novel photosystem II subunit and (ii) required for maintaining optimal photosystem II activity under adverse growth conditions [1]. Number of members: 17

10 [1] Medline: 94298765. The chloroplast ycf8 open reading frame encodes a photosystem II polypeptide which maintains photosynthetic activity under adverse growth conditions. Monod C, Takahashi Y, Goldschmidt-Clermont M, Rochaix JD; *EMBO J* 1994;13:2747-2754.

15 997. PSI_8. PHOTOSYSTEM I REACTION CENTRE SUBUNIT VIII. Synonym(s)PSI-I. Gene name(s)PSAI. From *Hordeum vulgare* (Barley). Encoded on Chloroplast. Taxonomy Eukaryota; Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; *Hordeum*.

20 MAY HELP IN THE ORGANIZATION OF THE PSAI SUBUNIT. BELONGS TO THE PSAI FAMILY.

[1] SEQUENCE FROM N.A. MEDLINE; 90036933. Scheller H.V., Okkels J.S., Hoej P.B., Svendsen I., Roepstorff P., Moeller B.L.; "The primary structure of a 4.0-kDa photosystem I polypeptide encoded by the chloroplast *psaI* gene."; *J. Biol. Chem.* 264:18402-18406(1989).

25

998. PSI_PsaJ: Photosystem I reaction centre subunit IX / PsaJ

This family consists of the photosystem I reaction centre subunit IX or PsaJ from various organisms including *Synechocystis* sp. (strain pcc 6803), *Pinus thunbergii* (green pine) and *Zea mays* (maize). PsaJ Swiss:P19443 is a small 4.4kDa, chloroplastal encoded, hydrophobic subunit of the photosystem I reaction complex its function is not yet fully understood [1].

30

PsaJ can be cross-linked to PsaF Swiss:P12356 and has a single predicted transmembrane domain it has a proposed role in maintaing PsaF in the correct orientation to allow for fast electron transfer from soluble donor proteins to P700+ [1]. Number of members: 18

[1] Medline: 99238330. A large fraction of PsaF is nonfunctional in photosystem I complexes lacking the PsaJ subunit. Fischer N, Boudreau E, Hippler M, Drepper F, Haehnel W, Rochaix JD; Biochemistry 1999;38:5546-5552.

5 [2] Medline: 93252282. Genes encoding eleven subunits of photosystem I from the thermophilic cyanobacterium *Synechococcus* sp. Muhlenhoff U, Haehnel W, Witt H, Herrmann RG; Gene 1993;127:71-78.

999. PSII. Protein namePHOTOSYSTEM II P680 CHLOROPHYLL A APOPROTEIN.

10 Synonym(s)CP-47 PROTEIN. Gene name(s)PSBB. From *Hordeum vulgare* (Barley), Encoded on Chloroplast. Taxonomy Eukaryota; Viridiplantae; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; *Hordeum*.

FUNCTION: THIS PROTEIN CONJUGATES WITH CHLOROPHYLL & CATALYZES THE PRIMARY LIGHT-INDUCED PHOTOCHEMICAL PROCESSES OF PHOTOSYSTEM II. SUBCELLULAR LOCATION: CHLOROPLAST THYLAKOID MEMBRANE. SIMILARITY: BELONGS TO THE PSBB / PSBC FAMILY.

15 [1] SEQUENCE FROM N.A. STRAIN=CV. SABARLIS; MEDLINE; 89240047. Andreeva A.V., Buryakova A.A., Reverdatto S.V., Chakhmakhcheva O.G., Efimov V.A.; "Nucleotide sequence of the 5.2 kbp barley chloroplast DNA fragment, containing psbB-psbH-petB-petD gene cluster."; Nucleic Acids Res. 17:2859-2860(1989).

20 [2] SEQUENCE FROM N.A. STRAIN=CV. SABARLIS; MEDLINE; 92207253. Efimov V.A., Andreeva A.V., Reverdatto S.V., Chakhmakhcheva O.G.; "Photosystem II of rye. Nucleotide sequence of the psbB, psbC, psbE, psbF, psbH genes of rye and chloroplast DNA regions adjacent to them."; Bioorg. Khim. 17:1369-1385(1991).

25 [3] SEQUENCE OF 411-420. Hinz U.G.; "Isolation of the photosystem II reaction center complex from barley. Characterization by circular dichroism spectroscopy and amino acid sequencing."; Carlsberg Res. Commun. 50:285-298(1985).

30 1000. QRPTase. Quinolate phosphoribosyl transferase.
Quinolate phosphoribosyl transferase (QRPTase) or nicotinate-nucleotide pyrophosphorylase EC:2.4.2.19 is involved in the de novo synthesis of NAD in both prokaryotes and eukaryotes. It catalyses the reaction of quinolinic acid with 5-

phosphoribosyl-1-pyrophosphate (PRPP) in the presence of Mg²⁺ to give rise to nicotinic acid mononucleotide (NaMN), pyrophosphate and carbon dioxide [1,2]. Number of members: 26.

- 5 [1]Medline: 97169443. A new function for a common fold: the crystal structure of quinolinic acid phosphoribosyltransferase. Eads JC, Ozturk D, Wexler TB, Grubmeyer C, Sacchettini JC; Structure 1997;5:47-58.
- [2]Medline: 96139309. The sequencing expression, purification, and steady-state kinetic analysis of quinolinate phosphoribosyl transferase from Escherichia coli. Bhatia R, Calvo
- 10 KC; Arch Biochem Biophys 1996;325:270-278.

1001. R3H domain

The name of the R3H domain comes from the characteristic spacing of the most conserved arginine and histidine residues. The function of the domain is predicted to be binding ssDNA. Number of members: 28

- [1]Medline: 99003905 The R3H motif: a domain that binds single-stranded nucleic acids. Grishin NV; Trends Biochem Sci 1998;23:329-330.

20 1002. recF protein signatures (RecF)

The prokaryotic protein recF [1,2] is a single-stranded DNA-binding protein which also probably binds ATP. RecF is involved in DNA metabolism; it is required for recombinational DNA repair and for induction of the SOS response. RecF is a protein of about 350 to 370

25 amino acid residues; there is a conserved ATP-binding site motif 'A' (P-loop) in the N-terminal section of the protein as well as two other conserved regions, one located in the central section, and the other in the C-terminal section. Signature patterns were derived from these two regions.

- 30 Consensus pattern [LIVM]-x(4)-[LIF]-x(6)-[LIF]-[LVF]-x-[GE]-[GSTAD]-[PA]- x(2)-R-R-x-[FYW]-[LIVMF]-D Sequences known to belong to this class detected by the pattern ALL.

Consensus pattern[LIVMFY](2)-x-D-x(2,3)-[SA]-[EH]-L-D-x(2)-[KRH]-x(3)-L Sequences known to belong to this class detected by the patternALL, except for T. palidum recF.

[1] Sandler S.J., Chackerian B., Li J.T., Clark A.J. Nucleic Acids Res. 20:839-845(1992).

5 [2] Alonso J.C., Fisher L.M.; Mol. Gen. Genet. 246:680-686(1995).

1003. RibD C-terminal domain (RibD_C)

10 The function of this domain is not known, but it is thought to be involved in riboflavin biosynthesis. This domain is found in the C terminus of RibD/RibG Swiss:P25539, in combination with dCMP_cyt_deam, as well as in isolation in some archaebacterial proteins Swiss:P95872.

Number of members: 21

15 1004. Ribosomal protein L16 signatures (Ribosomal_L16)

Ribosomal protein L16 is one of the proteins from the large ribosomal subunit. In Escherichia coli, L16 is known to bind directly the 23S rRNA and to be located at the A site of the peptidyltransferase center. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups:

- Eubacterial L16.
- Algal and plant chloroplast L16.
- Cyanelle L16.
- Plant mitochondrial L16.

25 L16 is a protein of 133 to 185 amino-acid residues. As signature patterns, we selected two conserved regions in the central section of these proteins.

Consensus pattern [KR](2)-x-[GSAC]-[KQVA]-[LIVM]-W-[LIVM]-[KR]-[LIVM]-[LFY]-[AP] Sequences known to belong to this class detected by the pattern ALL.

30

Consensus patternR-M-G-x-[GR]-K-G-x(4)-[FWKR] Sequences known to belong to this class detected by the patternALL.

[1] Otaka E., Hashimoto T., Mizuta K., Suzuki K. Protein Seq. Data Anal. 5:301-313(1993).

1005. Ribosomal protein L32e signature (Ribosomal_L32E)

A number of eukaryotic and archaebacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:

- Mammalian L32 [1].
- Drosophila RP49 [2].
- Trichoderma harzianum L32 [3].

- Yeast L32e (YBL092w).

- Archaeobacterial L32e [4].

These proteins have 135 to 240 amino-acid residues. As a signature pattern, a stretch of about 20 residues located in the N-terminal part of these proteins was selected.

Consensus pattern F-x-R-x(4)-[KR]-x(2)-[KR]-[LIVMF]-x(3,5)-W-R-[KR]-x(2)-G Sequences known to belong to this class detected by the pattern ALL.

[1] Jacks C.M., Powaser C.B., Hackett P.B. Gene 74:565-570(1988).

[2] Aguade M. Mol. Biol. Evol. 5:433-441(1988).

[3] Lora J.M., Garcia I., Benitez T., Llobell A., Pintor-Toro J.A. Nucleic Acids Res. 21:3319-3319(1993).

[4] Arndt E., Scholzen T., Kroemer W., Hatakeyama T., Kimura M. Biochimie 73:657-668(1991).

1006. (Ribosomal_S3) Ribosomal protein S3 signature

PROSITE: PDOC00474. PROSITE cross-reference(s) PS00548; RIBOSOMAL_S3

Ribosomal protein S3 is one of the proteins from the small ribosomal subunit. In Escherichia coli, S3 is known to be involved in the binding of initiator Met-tRNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups:

-Eubacterial S3.

-Algal and plant chloroplast S3.

-Cyanelle S3.

-Archaeobacterial S3.

-Plant mitochondrial S3.

-Vertebrate S3.

-Insect S3.

5 -Caenorhabditis elegans S3 (C23G10.3).

-Yeast S3 (Rp13).

S3 is a protein of 209 to 559 amino-acid residues. A conserved region located in the C-terminal section was selected as a signature pattern.

10 Consensus pattern[GSTA]-[KR]-x(6)-G-x-[LIVMT]-x(2)-[NQSCH]-x(1,3)-[LIVFCA]-x(3)-[LIV]-[DENQ]-x(7)-[LMT]-x(2)-G-x(2)-[GS]. Sequences known to belong to this class detected by the patternALL, except for some mitochondrial S3.

[1]Otake E., Hashimoto T., Mizuta K. Protein Seq. Data Anal. 5:285-300(1993).

1007. RimM - RimM

The RimM protein is essential for efficient processing of 16S rRNA [1]. The RimM protein was shown to have affinity for free ribosomal 30S subunits but not for 30S subunits in the 70S ribosomes [1]. Number of members: 14.

[1]Medline: 98083058. RimM and RbfA are essential for efficient processing of 16S rRNA in Escherichia coli. Bylund GO, Wipemo LC, Lundberg LA, Wikstrom PM; J Bacteriol 1998;180:73-82.

25 1008. RNA_pol_A - RNA polymerase alpha subunit

-!- RNA polymerases catalyse the DNA dependent polymerisation of RNA. Prokaryotes contain a single RNA polymerase compared to three in eukaryotes (not including mitochondrial and chloroplast polymerases).

-!- Members of this family include: A subunit from eukaryotes, gamma subunit from cyanobacteria, beta' subunit from eubacteria, A' subunit from archaeobacteria, B'' from chloroplasts. Number of members: 139.

30

[1]Medline: 97066998. Structural modules of the large subunits of RNA polymerase. Introducing archaebacterial and chloroplast split sites in the beta and beta' subunits of Escherichia coli RNA polymerase. Severinov K, Mustaev A, Kukarin A, Muzzin O, Bass I, Darst SA, Goldfarb A; J Biol Chem 1996;271:27969-27974.

5

1009. RuBisCO_large - Ribulose biphosphate carboxylase large chain active site
PROSITE: PDOC00142; PROSITE cross-reference(s) PS00157; RUBISCO_LARGE

Ribulose biphosphate carboxylase (EC 4.1.1.39) (RuBisCO) [1,2] catalyzes the initial step in Calvin's reductive pentose phosphate cycle in plants as well as purple and green bacteria. It consists of a large catalytic unit and a small subunit of undetermined function. In plants, the large subunit is coded by the chloroplastic genome while the small subunit is encoded in the nuclear genome. Molecular activation of RuBisCO by CO₂ involves the formation of a carbamate with the epsilon-amino group of a conserved lysine residue. This carbamate is stabilized by a magnesium ion. One of the ligands of the magnesium ion is an aspartic acid residue close to the active site lysine [3]. A pattern was developed which includes both the active site residue and the metal ligand, and which is specific to RuBisCO large chains.

10

15

20

Consensus patternG-x-[DN]-F-x-K-x-D-E [K is the active site residue] [The second D is a magnesium ligand]. Sequences known to belong to this class detected by the patternALL, except for Cheilopleuria bicuspidis RuBisCO.

[1]Miziorko H.M., Lorimer G.H. Annu. Rev. Biochem. 52:507-535(1983).

[2]Akazawa T., Takabe T., Kobayashi H. Trends Biochem. Sci. 9:380-383(1984).

[3]Andersson I., Knight S., Schneider G., Lindqvist Y., Lundqvist T., Branden C.-I., Lorimer G.H. Nature 337:229-234(1989).

25

1010. Rve - Integrase core domain

Integrase mediates integration of a DNA copy of the viral genome into the host chromosome. Integrase is composed of three domains. The amino-terminal domain is a zinc binding domain Integrase_Zn. This domain is the central catalytic domain. The carboxyl terminal domain that is a non-specific DNA binding domain integrase. The catalytic domain acts as an endonuclease when two nucleotides are removed from the 3' ends of the blunt-ended viral

30

DNA made by reverse transcription. This domain also catalyses the DNA strand transfer reaction of the 3' ends of the viral DNA to the 5' ends of the integration site [1]. Number of members: 694.

- 5 [1]Medline: 95099322. Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. Dyda F, Hickman AB, Jenkins TM, Engelman A, Craigie R, Davies DR; Science 1994;266:1981-1986.

1011. (SBP_bac_3) Bacterial extracellular solute-binding proteins, family 3 signature

10 PROSITE: PDOC00798. PROSITE cross-reference(s) PS01039; SBP_BACTERIAL_3

Bacterial high affinity transport systems are involved in active transport of solutes across the cytoplasmic membrane. The protein components of these traffic systems include one or two transmembrane protein components, one or two membrane-associated ATP-binding proteins (ABC transporters; see <PDOC00185>) and a high affinity periplasmic solute-binding protein. The later are thought to bind the substrate in the vicinity of the inner membrane, and to transfer it to a complex of inner membrane proteins for concentration into the cytoplasm.

In gram-positive bacteria which are surrounded by a single membrane and have therefore no periplasmic region the equivalent proteins are bound to the membrane via an N-terminal lipid anchor. These homolog proteins do not play an integral role in the transport process per se, but probably serve as receptors to trigger or initiate translocation of the solute through the membrane by binding to external sites of the integral membrane proteins of the efflux system.

In addition at least some solute-binding proteins function in the initiation of sensory transduction pathways.

On the basis of sequence similarities, the vast majority of these solute-binding proteins can be grouped [1] into eight families of clusters, which generally correlate with the nature of the solute bound.

Family 3 groups together specific amino acids and opine-binding periplasmic proteins and a periplasmic homolog with catalytic activity:

-Histidine-binding protein (gene hisJ) of Escherichia coli and related bacteria. An homologous lipoprotein exists in Neisseria gonorrhoeae.

812

-Lysine/arginine/ornithine-binding proteins (LAO) (gene argT) of Escherichia coli and related bacteria are involved in the same transport system than hisJ. Both solute-binding proteins interact with a common membrane-bound receptor hisP of the binding protein dependent transport system HisQMP.

5 -Glutamine-binding proteins (gene glnH) of Escherichia coli and Bacillus stearothermophilus.

-Glutamate-binding protein (gene gluB) of Corynebacterium glutamicum.

-Arginine-binding proteins artI and artJ of Escherichia coli.

-Nopaline-binding protein (gene nocT) from Agrobacterium tumefaciens.

10 -Octopine-binding protein (gene occT) from Agrobacterium tumefaciens.

-Major cell-binding factor (CBF1) (gene: peb1A) from Campylobacter jejuni.

-Bacteroides nodosus protein aabA.

-Cyclohexadienyl/arogenate dehydratase of Pseudomonas aeruginosa, a periplasmic enzyme which forms an alternative pathway for phenylalanine biosynthesis.

15 -Escherichia coli protein fliY.

-Vibrio harveyi protein patH.

-Escherichia coli hypothetical protein ydhW.

-Bacillus subtilis hypothetical protein yckB.

-Bacillus subtilis hypothetical protein yckK.

20

The signature pattern is located near the N-terminus of the mature proteins.

Consensus pattern G-[FYIL]-[DE]-[LIVMT]-[DE]-[LIVMF]-x(3)-[LIVMA]-[VAGC]-x(2)-[LIVMAGN]

Sequences known to belong to this class detected by the pattern ALL.

25

[1] Tam R., Saier M.H. Jr. Microbiol. Rev. 57:320-346(1993).

1012. Sec7 - Sec7 domain

The Sec7 domain is a guanine-nucleotide-exchange-factor (GEF) for the arf family [2].

30

Number of members: 32.

[1]Medline: 98169075. Structure of the Sec7 domain of the Arf exchange factor. ARNO. Cherfils J, Menetrey J, Mathieu M, Le Bras G, Robineau S, Beraud-Dufour S, Antonny B, Chardin P; Nature 1998;392:101-105.

[2]Medline: 97100951. A human exchange factor for ARF contains Sec7- and pleckstrin-homology domains. Chardin P, Paris S, Antonny B, Robineau S, Beraud-Dufour S, Jackson CL, Chabre M. Nature 1996;384:481-484.

1013. SecA_protein. SecA protein, amino terminal region

SecA protein binds to the plasma membrane where it interacts with proOmpA to support translocation of proOmpA through the membrane. SecA protein achieves this translocation, in association with SecY protein, in an ATP dependent manner. SecA possesses the ATPase activity. The carboxyl terminus has similarity with the helicase carboxyl terminus. See Ribosomal_L5. Number of members: 45.

[1]Medline: 98309858. Amino-terminal region of SecA is involved in the function of SecG for protein translocation into Escherichia coli membrane vesicles. Mori H, Sugiyama H, Yamanaka M, Sato K, Tagaya M, Mizushima S; J Biochem (Tokyo) 1998;124:122-129.

[2]Medline: 89251629. SecA protein hydrolyzes ATP and is an essential component of the protein translocation ATPase of Escherichia coli. Lill R, Cunningham K, Brundage LA, Ito K, Oliver D, Wickner W; EMBO J 1989;8:961-966.

1014. Seedstore_2S - 2S seed storage family

Members of this family are composed of two chains (both included in the alignment), these are co-translated and later cleaved. The two chains are disulphide linked together. Number of members: 27.

[1]Medline: 97121264. 1H NMR assignment and global fold of napin BnIb, a representative 2S albumin seed protein. Rico M, Bruix M, Gonzalez C, Monsalve RI, Rodriguez R; Biochemistry 1996;35:15672-15682.

1015. Smr - Smr domain

This family includes the Smr (Small MutS Related) proteins, and the C-terminal region of the MutS2 protein. It has been suggested that this domain interacts with the MutS1 Swiss:P23909

protein in the case of Smr proteins and with the N-terminal MutS related region of MutS2 Swiss:P94545 [1]. Number of members: 14.

[1]Medline: 10431172. Smr: a bacterial and eukaryotic homologue of the C-terminal region of the MutS2 family. Moreira D, Philippe H; Trends Biochem Sci 1999;24:298-300.

1016. (SSF) Sodium:solute symporter family signatures and profile

PROSITE: PDOC00429. PROSITE cross-reference(s)PS00456; NA_SOLUT_SYMP_1

PS00457; NA_SOLUT_SYMP_2 PS50283; NA_SOLUTE_SYMP_3

It has been shown [1,2] that integral membrane proteins that mediate the intake of a wide variety of molecules with the concomitant uptake of sodium ions (sodium symporters) can be grouped, on the basis of sequence and functional similarities into a number of distinct families. One of these families is known as the sodium:solute symporter family (SSF) and currently consists of the following proteins:

- Mammalian Na⁺/glucose co-transporter.
- Mammalian Na⁺/myo-inositol co-transporter.
- Mammalian Na⁺/nucleoside co-transporter.
- Mammalian Na⁺/neutral amino acid co-transporter.
- Escherichia coli Na⁺/proline symporter (gene putP).
- Escherichia coli Na⁺/pantothenate symporter (gene panF).
- Escherichia coli hypothetical protein yidK.
- Escherichia coli hypothetical protein yjcG.
- Bacillus subtilis hypothetical protein ywCA (ipa-31R).

These integral membrane proteins are predicted to comprise at least ten membrane spanning domains. Two conserved regions were selected as signature patterns; the first one is located in the fourth transmembrane region and the second one in a loop between two transmembrane regions in the C-terminal part of these proteins.

Consensus pattern[GS]-x(2)-[LIY]-x(3)-[LIVMFYWSTAG](10)-[LIY]-[TAV]-x(2)-G-G-[LMF]-x-[SAP]. Sequences known to belong to this class detected by the patternALL.

Consensus pattern[GAST]-[LIVM]-x(3)-[KR]-x(4)-G-A-x(2)-[GAS]-[LIVMGS]-[LIVMW]-[LIVMGAT]-G-x-[LIVMGA] Sequences known to belong to this class detected by the patternALL, except for E.coli yidK.

Note this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so.

- 5 [1]Reizer J., Reizer A., Saier M.H. Jr. Res. Microbiol. 141:1069-1072(1991).
[2]Reizer J., Reizer A., Saier M.H. Jr. Biochim. Biophys. Acta 1197:133-136(1994).

1017. SurE - Survival protein SurE

E. coli cells with the surE gene disrupted are found to survive poorly in stationary phase [1].

- 10 It is suggested that SurE may be involved in stress response. Yeast also contains a member of the family Swiss:P38254. Swiss:P30887 can complement a mutation in acid phosphatase, suggesting that members of this family could be phosphatases. Number of members: 17.

[1]Medline: 95014035. A new gene involved in stationary-phase survival located at 59 minutes on the Escherichia coli chromosome. Li C, Ichikawa JK, Ravetto JJ, Kuo HC, Fu JC, Clarke S; J Bacteriol 1994;176:6015-6022.

[2]Medline: 93046805. Complementation of Saccharomyces cerevisiae acid phosphatase mutation by a genomic sequence from the yeast Yarrowia lipolytica identifies a new phosphatase. Treton BY, Le Dall MT, Gaillardin CM; Curr Genet 1992;22:345-355.

1018. Synuclein - Synuclein

There are three types of synucleins in humans, these are called alpha, beta and gamma.

Alpha synuclein has been found mutated in families with autosomal dominant Parkinson's disease. A peptide of alpha synuclein has also been found in amyloid plaques in Alzheimer's patients. Number of members: 12.

[1]Medline: 98424410. The synuclein family. Lavedan C; Genome Res 1998;8:871-880.

1019. (T-box) T-box domain signatures

- 30 PROSITE: PDOC00972. PROSITE cross-reference(s) PS01283; TBOX_1 PS01264; TBOX_2

A number of eukaryotic DNA-binding proteins contain a domain of about 170 to 190 amino acids known as the T-box domain [1,2,3] and which probably binds DNA. The T-box

816

has first been found in the mice T locus (Brachyury) protein, a transcription factor involved in mesoderm differentiation. It has since been found in the following proteins:

- Vertebrate and invertebrate homologs of the T protein.
- Mammalian proteins TBX1 to TBX6.
- 5 -Mammalian protein TBR1 which is expressed specifically in brain.
- Xenopus laevis eomesodermin (eomes).
- Xenopus laevis Vegt (or Antipodean), a transcription factor that activates the expression of wnt-8, eomes and Brachyury.
- Chicken TbxT.
- 10 -Drosophila protein optomotor-blind (omb).
- Drosophila protein brachyenteron (byn) (also known as Trg), which is required for the specification of the hindgut and anal pads.
- Drosophila protein H15.
- Caenorhabditis elegans protein tbx-12.
- 15 -Caenorhabditis elegans hypothetical proteins F21H11.3, F40H6.4, T07C4.2, T07C4.6 and ZK177.10.

Two conserved regions were selected as signature patterns for the T-domain. The first region corresponds to the N-terminal of the domain and the second one to the central part.

Consensus pattern L-W-x(2)-[FC]-x(3,4)-[NT]-E-M-[LIV](2)-T-x(2)-G-[RG]-[KRO]

Sequences known to belong to this class detected by the patternALL, except for C.elegans ZK177.10.

Consensus pattern [LIVMYW]-H-[PADH]-[DEN]-[GS]-x(3)-G-x(2)-W-M-x(3)-[IVA]-x-F

Sequences known to belong to this class detected by the patternALL, except for C.elegans

25 tbx-12, ZK177.10 and Drosophila H15.

[1] Bollag R.J., Siegfried Z., Cebra-Thomas J.A., Garvey N., Davison E.M., Silver L.M. Nat. Genet. 7:383-389(1994).

[2] Agulnik S.I., Garvey N., Hancock S., Ruvinsky I., Chapman D.L., Agulnik I., Bollag R.J., Papaioannou V.E., Silver L.M. Genetics 144:249-254(1996).

[3] Papaioannou V.E. Trends Genet. 13:212-213(1997).

1020. Toprim - Toprim domain

This is a conserved region from DNA primase. This corresponds to the Toprim domain common to DnaG primases, topoisomerases, OLD family nucleases and RecR proteins [1]. Both DnaG motifs IV and V are present in the alignment, the DxD (V) motif may be involved in Mg²⁺ binding and mutations to the conserved glutamate (IV) completely abolish DnaG type primase activity [1]. DNA primase EC:2.7.7.6 is a nucleotidyltransferase it synthesizes the oligoribonucleotide primers required for DNA replication on the lagging strand of the replication fork; it can also prime the leading stand and has been implicated in cell division [2]. Number of members: 133.

- [1]Medline: 98391745. Toprim--a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. Aravind L, Leipe DD, Koonin EV; Nucleic Acids Res 1998;26:4205-4213.
- [2]Medline: 97368180. Cloning and analysis of the dnaG gene encoding Pseudomonas putida DNA primase. Szafranski P, Smith CL, Cantor CR; Biochim Biophys Acta 1997;1352:243-248.
- [3]Medline: 94124015. The Haemophilus influenzae dnaG sequence and conserved bacterial primase motifs. Versalovic J, Lupski JR; Gene 1993;136:281-286.

1021. TraB - TraB family

pAD1 is a hemolysin/bacteriocin plasmid originally identified in Enterococcus faecalis DS16. It encodes a mating response to a peptide sex pheromone, cAD1, secreted by recipient bacteria. Once the plasmid pAD1 is acquired, production of the pheromone ceases--a trait related in part to a determinant designated traB. However a related protein is found in C. elegans Swiss:Q94217, suggesting that members of the TraB family have some more general function. Number of members: 12.

- [1]Medline: 94302142. Characterization of the determinant (traB) encoding sex pheromone shutdown by the hemolysin/bacteriocin plasmid pAD1 in Enterococcus faecalis. An FY, Clewell DB; Plasmid 1994;31:215-221.

1022. (Transpo_mutator) Transposases, Mutator family, signature
PROSITE: PDOC00770. PROSITE cross-reference(s) PS01007;
TRANSPOSASE_MUTATOR

Autonomous mobile genetic elements such as transposon or insertion sequences (IS) encode an enzyme, called transposase, required for excising and inserting the mobile element. On the basis of sequence similarities, transposases can be grouped into various families. One of these families has been shown [1,2,3,E1] to consist of transposases from the following elements:

- Mutator from Maize.
- Is1201 from *Lactobacillus helveticus*.
- Is905 from *Lactococcus lactis*.
- Is1081 from *Mycobacterium bovis*.
- Is6120 from *Mycobacterium smegmatis*.
- Is406 from *Pseudomonas cepacia*.
- IsRm3 from *Rhizobium meliloti*.
- IsRm5 from *Rhizobium meliloti*.
- Is256 from *Staphylococcus aureus*.
- IsT2 from *Thiobacillus ferrooxidans*.

The maize Mutator transposase (MudrA) is a protein of 823 amino acids; the bacterial transposases listed above are proteins of 300 to 420 amino acids. These proteins contain a conserved domain of about 130 residues; a signature pattern was derived from the most conserved part of this domain.

Consensus pattern D-x(3)-G-[LIVMF]-x(6)-[STAV]-[LIVMFYW]-[PT]-x-[STAV]-x(2)-[QR]-x-C-x(2)-H. Sequences known to belong to this class detected by the pattern ALL.

[1]Eisen J.A., Benito M.-I., Walbot V. *Nucleic Acids Res.* 22:2634-2636(1994).

[2]Guilhot C., Gicquel B., Davies J., Martin C. *Mol. Microbiol.* 6:107-113(1992).

[3]Wood M.S., Byrne A., Lessie T.G. *Gene* 105:101-105(1991).

1023. Transposase_8 - Transposase

Transposase proteins are necessary for efficient DNA transposition. This family consists of various *E. coli* insertion elements and other bacterial transposases some of which are members of the IS3 family. Number of members: 58.

[1]Medline: 97324595. Genetic organization and transposition properties of IS511. D. A. Mullin, D. L. Zies, A. H. Mullin, N. Caballera & B. Ely; Mol Gen Genet 1997;254:456-463.
 [2]Medline: 97128810. The use of an improved transposon mutagenesis system for DNA sequencing leads to the characterization of a new insertion sequence of *Streptomyces lividans* 66. J. Fischer, H. Maier, P. Viell & J. Altenbuchner; Gene 1996;180:81-89.
 [3]Medline: 97074647. Identification and nucleotide sequence of *Rhizobium meliloti* insertion sequence ISRM6, a small transposable element that belongs to the IS3 family. S. Zekri & N. Toro; Gene 1996;175:43-48.

1024. tRNA_int_endo - tRNA intron endonuclease

Members of this family cleave pre tRNA at the 5' and 3' splice sites to release the intron
 EC:3.1.27.9. Number of members: 8.

[1]Medline: 97344075. Properties of *H. volcanii* tRNA intron endonuclease reveal a relationship between the archaeal and eucaryal tRNA intron processing systems. Kleman-Leyer K, Armbruster DW, Daniels CJ; Cell 1997;89:839-847.

1025. Urease - Urease signatures

PROSITE: PDOC00133PROSITE cross-reference(s) PS01120; UREASE_1 PS00145;
 UREASE_2

Urease (EC 3.5.1.5) is a nickel-binding enzyme that catalyzes the hydrolysis of urea to carbon dioxide and ammonia [1]. Historically, it was the first enzyme to be crystallized (in 1926). It is mainly found in plant seeds, microorganisms and invertebrates. In plants, urease is a hexamer of identical chains. In bacteria [2], it consists of either two or three different subunits (alpha, beta and gamma).

Urease binds two nickel ions per subunit; four histidine, an aspartate and a carbamated-lysine serve as ligands to these metals; an additional histidine is involved in the catalytic mechanism [3].

As signatures for this enzyme, a region that contains two histidine that bind one of the nickel ions and the region of the active site histidine was selected.

Consensus pattern T-[AY]-[GA]-[GAT]-[LIVM]-D-x-H-[LIVM]-H-x(3)-P [The two H's bind nickel].Sequences known to belong to this class detected by the patternALL.

Consensus pattern[LIVM](2)-[CT]-H-[HN]-L-x(3)-[LIVM]-x(2)-D-[LIVM]-x-F-A [H is the active site residue]. Sequences known to belong to this class detected by the patternALL.

[1]Takishima K., Suga T., Mamiya G. Eur. J. Biochem. 175:151-165(1988).

5 [2]Mobley H.L.T., Husinger R.P. Microbiol. Rev. 53:85-108(1989).

[3]Jabri E., Carr M.B., Hausinger R.P., Karplus P.A. Science 268:998-1004(1995).

1026. Urease_beta - Urease beta subunit.

This subunit is known as alpha in Heliobacter. Number of members: 35.

10

[1]Medline: 95273988. The crystal structure of urease from Klebsiella aerogenes. Jabri E, Carr MB, Hausinger RP, Karplus PA; Science 1995;268:998-1004.

1027. UvrD-helicase - UvrD/REP helicase

15

The Rep family helicases are composed of four structural domains. The Rep family function as dimers. REP helicases catalyse ATP dependent unwinding of double stranded DNA to single stranded DNA. Swiss:P23478, Swiss:P08394 have large insertions near to the carboxy-terminus relative to other members of the family. Number of members: 52.

20

[1] Medline: 97433075. Major domain swiveling revealed by the crystal structures of complexes of E. coli Rep helicase bound to single-stranded DNA and ADP. Korolev S, Hsieh J, Gauss GH, Lohman TM, Waksman G; Cell 1997;90:635-647.

1028. V-type ATPase 116kDa subunit family (V_ATPase_sub_a)

25

This family consists of the 116kDa V-type ATPase (vacuolar (H⁺)-ATPases) subunits, as well as V-type ATP synthase subunit i. The V-type ATPases family are proton pumps that acidify intracellular compartments in eukaryotic cells for example yeast central vacuoles, clathrin-coated and synaptic vesicles. They have important roles in membrane trafficking processes [1]. The 116kDa subunit (subunit a) in the V-type ATPase is part of the V0 functional domain responsible for proton transport. The a subunit is a transmembrane glycoprotein with multiple putative transmembrane helices. It has a hydrophilic amino terminal and a hydrophobic carboxy terminal [1,2]. It has roles in proton transport and

30

821

assembly of the V-type ATPase complex [1,2]. This subunit is encoded by two homologous gene in yeast VPH1 and STV1 [2].

Number of members: 27

- 5 [1] Forgac M; Medline: 99240666 "Structure and properties of the vacuolar (H⁺)-ATPases." J Biol Chem 1999;274:12951-12954.
[2] Forgac M; Medline: 99270697 "Structure and properties of the clathrin-coated vesicle and yeast vacuolar V-ATPases." J Bioenerg Biomembr 1999;31:57-65.

10 1029. Viral (Superfamily 1) RNA helicase (Viral_helicase1)

Number of members: 260

- 15 [1] Koonin EV, Dolja VV; Medline: 94094568 "Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences." Crit Rev Biochem Mol Biol 1993;28:375-430.

1030. Vesicular monoamine transporter (VMAT)

20 This family consists of various vesicular amine transporters with 12 transmembrane helices. These included vesicular acetylcholine transporters (VACHT) [3], and vesicular monoamine transporters (VMATs) [1,2] isoforms 1 adrenal and 2 brain (VMAT1 and VMAT2).

25 These proteins transport biogenic amines into synaptic vesicles or chromaffin granules [4]. VMATs pack monoamine neurotransmitters into secretory vesicles for regulated exocytotic release, they also protect against the parkinsonian neurotoxins MPP⁺ by transporting it into vesicles preventing it from acting on mitochondria [1].

30 Also in the family is C. elegans UNC-17 a putative vesicular acetylcholine transporter mutations in UNC-17 cause impaired neuromuscular function, giving rise to jerky or uncoordinated movement, [4].

Number of members: 15

- [1] Krantz DE, Peter D, Liu Y, Edwards RH; Medline: 97197857 "Phosphorylation of a vesicular monoamine transporter by casein kinase II." J Biol Chem 1997;272:6752-6759.
- [2] Erickson JD, Varoqui H, Schafer MK, Modi W, Diebler MF, Weihe E, Rand J, Eiden LE, Bonner TI, Usdin TB; Medline: 94350930 "Functional identification of a vesicular acetylcholine transporter and its expression from a 'cholinergic' gene locus." J Biol Chem 1994;269:21929-21932.
- [3] Erickson JD, Schafer MK, Bonner TI, Eiden LE, Weihe E; Medline: 96209876 "Distinct pharmacological properties and distribution in neurons and endocrine cells of two isoforms of the human vesicular monoamine transporter." Proc Natl Acad Sci U S A 1996;93:5166-5171.
- [4] Alfonso A, Grundahl K, Duerr JS, Han HP, Rand JB; Medline: 3342494 "The *Caenorhabditis elegans* unc-17 gene: a putative vesicular acetylcholine transporter." Science 1993;261:617-619.

1031. WW/rsp5/WWP domain signature and profile. Cross-reference(s): PS01159; WW_DOMAIN_1; PS50020; WW_DOMAIN_2

The WW domain [1-4,E1] (also known as rsp5 or WWP) has been originally discovered as a short conserved region in a number of unrelated proteins, among them dystrophin, the gene responsible for Duchenne muscular dystrophy. The domain, which spans about 35 residues, is repeated up to 4 times in some proteins. It has been shown [5] to bind proteins with particular proline-motifs, [AP]-P-P-[AP]-Y, and thus resembles somewhat SH3 domains. It appears to contain beta-strands grouped around four conserved aromatic positions; generally Trp. The name WW or WWP derives from the presence of these Trp as well as that of a conserved Pro. It is frequently associated with other domains typical for proteins in signal transduction processes.

Proteins containing the WW domain are listed below.

--Dystrophin, a multidomain cytoskeletal protein. Its longest alternatively spliced form consists of an N-terminal actin-binding domain, followed by 24 spectrin-like repeats, a cysteine-rich calcium-binding domain and a C-terminal globular domain. Dystrophin form tetramers and is thought to have multiple functions including involvement in membrane stability, transduction of contractile forces to the extracellular environment and organization

of membrane specialization. Mutations in the dystrophin gene lead to muscular dystrophy of Duchenne or Becker type. Dystrophin contains one WW domain C-terminal of the spectrin-repeats.

--Utrophin, a dystrophin-like protein of unknown function.

5 --Vertebrate YAP protein is a substrate of an unknown serine kinase. It binds to the SH3 domain of the Yes oncoprotein via a proline-rich region. This protein appears in alternatively spliced isoforms, containing either one or two WW domains [6].

--Mouse NEDD-4 plays a role in the embryonic development and differentiation of the central nervous system. It contains 3 WW modules followed by a HECT domain. The
10 human ortholog contains 4 WW domains, but the third WW domain is probably spliced resulting in an alternate NEDD-4 protein with only 3 WW modules [3].

--Yeast RSP5 is similar to NEDD-4 in its molecular organization. It contains an N-terminal C2 domain (see <PDOC00380>), followed by a histidine-rich region, 3 WW domains and a HECT domain.

15 --Rat FE65, a transcription-factor activator expressed preferentially in liver. The activator domain is located within the N-terminal 232 residues of FE65, which also contain the WW domain.

--Yeast ESS1/PTF1, a putative peptidyl prolyl cis-trans isomerase from family ppiC (see <PDOC00840>). A related protein, dodo (gene dod) exists in Drosophila and in mammals
20 (gene PIN1).

--Tobacco DB10 protein. The WW domain is located N-terminal to the region with similarity to ATP-dependent RNA helicases.

--IQGAP, a human GTPase activating protein acting on ras. It contains an N-terminal domain similar to fly muscle mp20 protein and a C-terminal ras GTPase activator domain.

25 --Yeast pre-mRNA processing protein PRP40, Caenorhabditis elegans ZK1098.1 and fission yeast SpAC13C5.02 are related proteins with similarity to MYO2-type myosin, each containing two WW-domains at the N-terminus.

--Caenorhabditis elegans hypothetical protein C38D4.5, which contains one WW module, a PH domain (see <PDOC50003>) and a C-terminal phosphatidylinositol 3-kinase domain.

30 --Yeast hypothetical protein YFL010c.

For the sensitive detection of WW domains, a profile was developed which spans the whole homology region as well as a pattern.

Description of pattern(s) and/or profile(s):

Consensus pattern W-x(9,11)-[VFY]-[FYW]-x(6,7)-[GSTNE]-[GSTQCR]-[FYW]-x(2)-P.

5

[1] Bork P., Sudol M. Trends Biochem. Sci. 19:531-533(1994).

[2] Andre B., Springael J.Y. Biochem. Biophys. Res. Commun. 205:1201-1205(1994).

[3] Hofmann K.O., Bucher P. FEBS Lett. 358:153-157(1995).

[4] Sudol M., Chen H.I., Bougeret C., Einbond A., Bork P. FEBS Lett. 369:67-71(1995).

10 [5] Chen H.I., Sudol M. Proc. Natl. Acad. Sci. U.S.A. 92:7819-7823(1995).

[6] Sudol M., Bork P., Einbond A., Kastury K., Druck T., Negrini M., Huebner K., Lehman D. J. Biol. Chem. 270:14733-14741(1995).

1032. XPA protein signatures. cross-reference(s): XPA_1 PROSITE PS00752;
PS00753;XPA_2.

15

Xeroderma pigmentosum (XP) [1] is a human autosomal recessive disease, characterized by a high incidence of sunlight-induced skin cancer. People's skin cells with this condition are hypersensitive to ultraviolet light, due to defects in the incision step of DNA excision repair. There are a minimum of seven genetic complementation groups involved in this pathway: XP-A to XP-G. XP-A is the most severe form of the disease and is due to defects in a 30 Kd nuclear protein called XPA (or XPAC) [2].

20

The sequence of the XPA protein is conserved from higher eukaryotes [3] to yeast (gene RAD14) [4]. XPA is a hydrophilic protein of 247 to 296 amino-acid residues which has a C4-type zinc finger motif in its central section.

25

Two signature were developed patterns for XPA proteins. The first corresponds to the zinc finger region, the second to a highly conserved region located some 12 residues after the zinc finger region.

30

Consensus pattern C-x-[DE]-C-x(3)-[LIVMF]-x(1,2)-D-x(2)-L-x(3)-F-x(4)-C-x(2)-C

Consensus pattern [LIVM](2)-T-[KR]-T-E-x-K-x-[DE]-Y-[LIVMF](2)-x-D-x-[DE]

[1] Tanaka K., Wood R.D. Trends Biochem. Sci. 19:83-86(1994).

[2] Miura N., Miyamoto I., Asahina H., Satokata I., Tanaka K., Okada Y. J. Biol. Chem. 266:19786-19789(1991).

5 [3] Shimamoto T., Kohno K., Tanaka K., Okada Y. Biochem. Biophys. Res. Commun. 181:1231-1237(1991).

[4] Bankmann M., Prakash L., Prakash S. Nature 355:555-558(1992).

1033. YCF9

10 This family consists of the hypothetical protein product of the YCF9 gene from chloroplasts and cyanobacteria. Number of members: 16

1034. (DUF15)

15 It is highly conserved between eubacteria and eukaryotes.

Number of members: 30

1035. Luminal portion of Cytochrome b559, alpha (gene psbE) subunit. (cytochr_b559a)

20 This family is the luminal portion of cytochrome b559 alpha chain, matches to this family should be accompanied by a match to the cytochr_b559 family also. The Prosite pattern pattern matches the transmembrane region of the cytochrome b559 alpha and beta subunits.

Number of members: 16

25

A. Asparaginase 2

30 Asparaginase II (L-asparagine aminohydrolase II) is an extracellular protein that may be associated with the cell wall and whose expression is affected by the availability of nitrogen. Asparaginase II catalyzes the reaction of L-Asparagine + H₂O = L-Aspartate + NH₃. As many leukemias have high requirements for aspartic acid, asparaginase II proteins are useful

as reagents for screening compounds for activity as leukemia chemotherapy products. Asparaginase II protein can also be over- or under-expressed to alter amino acid content in plant tissues or to modify nitrogen fixation and/or nitrogen metabolism in plants.

- 5 Ref: Bon et al. (1997) Appl Biochem Biotechnol 63-65: 203-12

B. Chloroa b-bind

Chlorophyll a-b binding proteins are located in the thylakoid membranes of the chloroplast and bind chlorophyll a and chlorophyll b, thereby triggering a chemical reaction (photosynthesis). These proteins are useful in controlling the rate, efficiency and/or output of photosynthesis. Overexpression of chlorophyll a-b binding proteins is expected to increase the rate of photosynthesis.

- 15 Ref: Leutwiler et al. (1986) Nucleic Acids Res 14: 4051-64
Brandt et al. (1992) Plant Mol Biol 19: 699-703

C. DMRL synthase

DMRL Synthase (6,7-Dimethyl-8-Ribityllumazine Synthase) catalyzes the last step in riboflavin (Vitamin B₂) synthesis, condensing 5-amino-6-(1'-D)-ribityl-amino-2,4(1H, 3H)-Pyrimidinedione with L-3,4-Dihydroxy-2-Butanone 4-Phosphate producing 6,7-Dimethyl-8-(1-D-Ribityl)Luminazine . The enzyme forms a homopentamer. Engineering of these proteins or those with homologous sequences/structures may allow control of the amounts of vitamin B₂ available in plants and/or accumulation of pigment, as well as altering reactions requiring hydrogen ion carriers/transmitters.

- Ref: Garcia-Ramirez et al. (1995) J Biol Chem **270**: 23801-7

D. E1_N

These proteins are ATP-dependent DNA helicases that are required for initiation of viral DNA replication. They form a complex with the viral E2 protein. The E1-E2 complex binds

to the replication origin that contains binding sites for both proteins. The majority of sequences known for this group of proteins are from various papillomaviruses, a type of double stranded DNA virus. In plants, the prototype double stranded DNA virus is Cauliflower Mosaic virus (CaMV). Manipulation of these proteins, especially to produce variant proteins that form non-productive complexes, enables production of plants that are resistant to infection by double stranded DNA viruses.

Ref: Yang et al. (1993) PNAS USA **90**: 5086-90

Ustav and Stenlund (1991) EMBO J **10**: 449-57

Callaway et al. (1996) Mol Plant Microbe Interact 9: 810-8

E. EF1_G

Elongation Factor-1 is composed of four subunits: alpha, beta, delta and gamma. Gamma subunits are presumed to play a role in anchoring the complex to other cellular components. Studies of EF-1 genes in plants suggests that different forms of the EF-1 subunits may be expressed in particular organs or in response to stress. Manipulation of the activity of these proteins, either by altered expression level or by structural mutation, may result in the accumulation of a particular protein in a chosen organ or allow production of particular proteins during stress conditions.

Ref: Kinzy et al. (1994) NAR 22: 2703-7

Dunn et al. (1993) Plant Mol Biol 23: 221-5

Aguilar et al. (1991) Plant Mol Biol 17: 351-60

F. ENV_polyprotein

This family comprises the envelope or coat proteins known from a number of different retroviruses. In mammalian species, retroviruses are responsible for diseases such as leukemia and HIV. In plants, retroviruses are known in both monocot (e.g. Zeon-1) and dicot (e.g. Arabidopsis and tobacco) species and have been shown to induce mutant alleles at new loci. Engineering of plant ENV proteins may allow mobilization or targeting of endogenous

or introduced retroviruses, in essence generating a new method for mutant production, gene tagging and the like.

Ref: Mamoun et al (1990) J Virol 64: 4180-8
5 Grandbastien et al. (1989) Nature 337: 376-80
Wright and Voytas (1998) Genetics 149: 703-15

G. Glycosyl_hydr9

10 Proteins having this domain (previously known as the glycosyl hydrolase family 5 domain) catalyze the endohydrolysis of 1,4- β -D-glucosidic linkages in cellulose. Numerous plant proteins with this domain exist and are expressed in an organ specific manner. They are involved in the fruit ripening process, in cell elongation and plant reproduction. Modulation
15 of the activity of these proteins, either by over- or under-expression or by mutation of the polypeptide, could be used to affect post-harvest physiology (e.g. rate of ripening) or for engineering reproductive sterility.

Ref: Giorda et al. (1990) Biochemistry 29: 7264-9
20 Tucker et al. (1988) Plant Physiol 88: 1257-62
Shani et al. (1997) 43: 837-42
Milligan and Gasser (1995) Plant Mol Biol 28: 691-711

H. Glycosyl_hydr14

25 The β -amylases (family 14 of glycosyl hydrolases) catalyze the hydrolysis of 1,4- α -glucosidic linkages in polysaccharides and remove successive maltose units from the non-reducing ends of the chains. Mutants of β -amylase in Arabidopsis exhibited altered degradation of starch throughout the diurnal cycle. In addition, the mutant phenotypes
30 indicated that these enzymes not only affect carbohydrate metabolism/catabolism, but also influence the amount of pigment stored within particular cells. Manipulation of the β -amylase genes enables control of plant pigmentation (for example, fibre pigment in cotton) as well as carbohydrate synthesis and degradation.

Ref: Zeeman et al. (1998) Plant J 15: 357-65
Hirano and Nakamura (1997) Plant Physiol 114: 5675-82
Kitamoto et al. (1988) J Bacteriol 170: 5848-54

5

I. Glycosyl_hydr15

Glycosyl hydrolases from family 15 (such as 1,4-Alpha-D-Glucan glucohydrolase,) catalyze the hydrolysis of terminal 1,4-linked alpha-D-glucose residues successively from the non-reducing ends of the chains resulting in the release of β -D-Glucose. In plants these proteins have been tied to the mobilization of the xyloglucan stored in the cotyledonary cell walls. Proteins such as these could be varied to affect the rate of plant growth (for example during germination), storage and/or use of glucose and other sugars by plant tissues and alteration of the properties, such as elasticity, of plant cell walls.

Ref: Crombie et al. (1998) Plant J 15: 27-38
Hata et al. (1991) Agric Biol Chem 55: 941-9

J. Glycosyl_hydr20

Members of the family 20 glycosyl hydrolases catalyze the hydrolysis of terminal non-reducing N-acetyl-D-hexosamine residues in N-acetyl- β -D-hexosaminides. N-acetyl- β -glucosaminidase belongs to this family and exists in several different forms (consisting of various combinations of alpha and beta chains) depending on the organism. Family 20 glycosyl hydrolases have been implicated in lysosomal storage diseases (such as Sandhoff disease) and glycogen storage disease in humans. These types of proteins are also responsible for the hydrolysis of chitin. In plants, these proteins could be useful in controlling carbohydrate catabolism, thereby influencing the amount of sugars available for storage and/or use in other metabolic pathways. In addition, it is possible that such proteins could be used to engineer an endogenous insect protection mechanism, e.g. by secretion of a chitin-hydrolyzing composition by the plant.

Ref: Graham et al (1988) J Biol Chem 263: 16823-9
O'Dowd et al. (1988) Biochemistry 27: 5216-26

K. HMG box

5

The HMG box is a novel type of DNA-binding domain found in a diverse group of proteins. Numerous plant proteins contain this domain, such as the HMG1/2-like proteins. The expression of some of these HMG proteins appears to be regulated by circadian rhythms and in a light dependent manner, occurring at higher levels in roots, for example and lower levels in light-grown tissues such as cotyledons. Generally, HMG proteins are thought to influence transcription regulation. In plants, HMGs are believed to have a role in maintaining patterns of circadian-regulated expression for other genes, suggesting that these proteins could be exploited to control growth and development.

10

Ref: Laudet et al. (1993) Nucleic Acids Res 21: 2493-501
Zheng et al. (1993) Plant Mol Biol 23: 813-23
Grasser et al. (1993) Plant Mol Biol 23: 619-25

15

L. IL2

20

Interleukin-2 (IL-2) is produced in mammals by T cells in response to antigenic or mitogenic stimulation and is crucial for proper regulation and functioning of the immune response. IL-2 is capable of stimulating B cells, monocytes, lymphokine-activated killer cells, natural killer cells and glioma cells. Plant extracts have also been shown to stimulate the immune system (for example, mistletoe therapy for human cancer). It is known that IL-2 is involved in feedback inhibition pathways that impact the inflammatory response as well as the growth inhibition of tumor reactive T cells. Plant proteins containing IL-2-like sequences are useful as immunity-based therapeutics, acting in a manner similar to IL-2 in mammals.

25

Ref: Heike et al. (1997) Scand J Immunol 45: 221-6
Ariel et al. (1998) J Immunol 161: 2465-72
Schink (1997) Anticancer Drugs 8 Suppl 1: S47-51

30

M. Oxidored_FMN

NADPH dehydrogenases catalyze the reaction $\text{NADPH} + \text{acceptor} = \text{NADP}(+) + \text{reduced acceptor}$. One member of this family is yeast “old yellow enzyme” (OYE) and is thought to be involved in oxylipin metabolism. A second yeast family member is a protein that binds estrogen binding protein (EBP) in addition to exhibiting oxidoreductase activity. An Arabidopsis homolog to OYE has been described and estrogen binding proteins in plants have been reported. Plant proteins from this class have the potential to be used to modify lipid metabolism/catabolism. These proteins may also have use as therapeutics for breast and prostate cancer, and other abnormal growth in steroid-sensitive tissues.

Ref: Baker et al. (1998) Proc Soc Exp Biol Med 217: 317-21
Schaller and Weiler (1997) J Biol Chem 272: 28066-72
Mandani et al. (1994) PNAS USA 91: 922-6

N. Oxidored_q2

The NADH-plastoquinone oxidoreductases catalyze the reaction $\text{NADH} + \text{plastoquinone} = \text{NAD}(+) + \text{plastoquinol}$. In plants these reactions occur in the chloroplast and are believed to participate in a chloroplast respiratory system. Here, the NDH complex is postulated to act as a valve to remove excess reduction equivalents in the chloroplasts. Manipulation of these proteins may improve the rate or efficiency of photosynthesis.

Ref: Burrows et al. (1998) EMBO J 17: 868-76
Kofer et al (1998) Mol Gen Genet 258: 166-73
Maier et al. (1995) J Mol Biol 251: 614-28

O. PABP

Polyadenylate binding proteins bind the poly (A) tail of mRNA. Plants, as exemplified by Arabidopsis, contain numerous PABP genes that are expressed in an organ-specific manner. For example, PABP2 is functional in roots and shoots, while PABP5 is expressed predominantly in immature flowers. The PABP proteins are implicated in numerous aspects

of posttranscriptional regulation including mRNA turnover and translational initiation. Control of activity of PABP proteins provides the ability to control the expression of various genes in particular organs during development.

- 5 Ref: Hilson et al (1993) Plant Physiol 103: 525-33
 Belostotsky and Meagher (1993) PNAS USA 90: 6686-90

P. Parvo coat

- 10 Parvoviruses are linear single-stranded DNA viruses that are encapsulated by three capsid proteins. Plants are susceptible to infection by single stranded DNA viruses such as Maize streak virus (MSV) and various Gemini viruses. The coat proteins in these plant viruses are critical to the virus life cycle within the plant. For example, the coat protein of MSV is thought to be involved in intra- and inter-cellular movement within the plant. Engineering of
15 proteins having similarity to parvoviral coat proteins, especially to produce proteins that interfere with maturation of the virus particle, enables the production of plants having better resistance to natural plant single-stranded DNA viruses.

- Ref: Liu et al. (1997) J Gen Virol 78: 1265-70
20 Rohde et al. (1990) Virology 176: 648-51

Q. Pkinase_C

- Plant serine/threonine protein kinases possessing this domain are expressed in all tissues and
25 are known to undergo serine-specific autophosphorylation and specifically phosphorylate two ribosomal proteins, P14 and P16. During development, these proteins predominate during high metabolic activity in growing buds, root tips, leaf margins and germinating seeds. They are thought to be involved in the control of plant growth and development. In addition, two genes encoding proteins from this family have been described that help plant cells adapt
30 during cold or high salt stresses. Consequently, engineering Pkinase C proteins provides a way to control general growth/development of the plant as well as a means to provide endogenous protection against environmental stresses.

Ref: Zhang et al. (1994) J Biol Chem 269: 17586-92

Mizoguchi et al. (1995) FEBS Lett 358: 199-204

R. REV

5

The REV proteins act post-transcriptionally to relieve negative repression of GAG and ENV production in retroviruses such as Human Immunodeficiency Virus type I (HIV-1). Plants contain retrovirus-like viruses such as pararetroviruses and retrotransposons (i.e. transposons having long terminal repeats). Plant retrotransposons in particular have been used to create mutations at various loci, thereby permitting gene isolation, gene tagging and the like. Manipulation of plant REV proteins enables control of transposition frequencies of corresponding transposable elements and provides a new tool for genetic engineering of plants.

10

Ref: Sodroski et al. (1986) Nature 321: 412-7
Franchini et al. (1989) PNAS USA 86: 2433-7
Marquet et al. (1995) 77: 113-24
Grandbastien et al. (1989) Nature 337: 376-80
Wright and Voytas (1998) Genetics 149: 703-15

15

20

S. RuBisCo small

Ribulose 1,5-bisphosphate carboxylase/oxygenase (RuBisCo) catalyzes the initial step in the C3 photosynthetic carbon reduction cycle, adding carbon dioxide to D-ribulose 1,5-bisphosphate to form two molecules of 3-phospho-D-glycerate. RuBisCo is comprised of two subunits, one large which is synthesized in the chloroplast, and one small which is synthesized in the cytoplasm and then transported in to the chloroplast. The expression of the small subunit of RuBisCo is light regulated. Manipulation of these proteins could increase the efficiency of photosynthesis or allow alterations in developmental timing.

25

30

Ref: Giuliano et al. (1988) PNAS USA 85: 7089-93
Dedonder et al. (1993) Plant Physiol 101: 801-8

T. Sialyltransf

Members of the CMP-N-acetylneuraminate- β -galactosamide- α -2,3-sialyltransferase family catalyze the following reaction:

5 CMP-N-acetylneuraminate + β -D-galactosyl-1,3-N-acetyl- α -D-galactosaminyl-R = CMP + α -N-acetylneraminyl-2,3- β -D-galactosyl-1,3-N-acetyl- α -D-galactosaminyl-R. These proteins are thought to be responsible for the synthesis of the sequence neurac- α -2,3-gal- β -1,3-galnac- found on sugar chains)-linked to threonine or serine and also as a terminal sequence on certain gangliosides in mammalian cells. In plants, glycosyltransferases in the Golgi apparatus synthesize cell wall polysaccharides and elaborate the complex glycans of glycoproteins. Engineering of plant sialyltransferases allows targeting of proteins to particular cellular locations or enables the making of changes in cell wall structure.

Ref: Wee et al. (1998) Plant Cell 10: 1759-68

Lee et al. (1994) J Biol Chem 269: 10028-33

Kitagawa and Paulson (1994) J Biol Chem 269: 1394-401

U. Signal

20 Many plant proteins in this family contain sequences similar to those found in both components of the prokaryotic family of signal transducers known as the two-component systems. This suggests that activation may require a transfer of a phosphate group between the transmitter domain and the receiver domain. One family member in Arabidopsis appears to be involved in ethylene (a plant hormone) signal transduction. Other proteins in this family appear to be involved in the regulation of gene transcription under conditions of environmental stress. Signal proteins can be exploited to affect plant growth and development and/or control plant responses to stress conditions such as cold, nutrient availability, etc.

Ref: Chang et al. (1993) Science 262: 539-44

30 Nagaya et al. (1993) Gene 131: 119-124

Gottfert et al. (1990) PNAS USA 87: 2680-4

V. vMSA

vMSA proteins are major surface antigens presenting on the envelope of various retroviruses. Surface antigens of retroviruses are often involved in tropism of the virus.

5 Plants contain retrovirus-like viruses such as pararetroviruses and retrotransposons (i.e. transposons having long terminal repeats). Plant retrotransposons in particular have been used to create mutants at various loci, thereby permitting gene isolation, gene tagging and the like. Manipulation of plant vMSA proteins enables control of tropism of plant retroviruses that might be used for genetic engineering tools, thus enabling targeting of the virus to
10 particular species and/or tissues of plants.

Ref: Okamoto et al. (1988) J Gen Virol 69: 2575-83

Grandbastien et al. (1989) Nature 337: 376-80

Wright and Voytas (1998) Genetics 149: 703-15

W. zf-CCCH

This family of proteins is defined by having two CX(8)CX(5)CX(3)H-type zinc finger domains. These proteins cover a broad range of functions. For example, the COP1 protein acts as a repressor of photomorphogenesis in darkness; light stimuli abolish this suppressive action. In addition, COP1 protein can function as a negative transcriptional regulator capable of direct interaction with components of the G-protein signaling pathway. As a second example, a zf-CCCH protein identified in Arabidopsis appears to be involved in the resistance to DNA damage induced by UV light and chemical DNA-damaging agents.

20 Overexpression of this class of proteins permits production of plants that are better suited to adverse environments. Manipulation of expression of zf-CCCH proteins functioning as transcriptional regulators, such as COP1, enables manipulation of some signal transduction pathways.

30 Ref: Pang et al. (1993) Nucleic Acids Res 21: 1647-53

Deng et al. (1992) Cell 71: 791-801

X. zf-RanBP

Proteins falling within this category contain many X-X-F-G and X-F-X-F-G repeats, and may contain RANBP1-like or PPIase domains. Plant proteins having domains similar to these include PAS1 and GMSTI. PAS1 has been shown to have dramatic developmental affects that appear to be correlated with both cell division and cell wall elongation. GMSTI has high identity to the yeast STI stress-inducible gene and has been shown to be heat inducible. Proteins such as these may be useful for controlling growth and form of development.

Ref: Vittorioso et al. (1998) Mol Cell Biol 18: 3034-43
Hernandez Torres et al. (1995) 27: 1221-6

Y. Peptidase M48.

Proteins belonging to this peptidase family are metalloproteases that bind zinc as a cofactor and are located in the membranes of the endoplasmic reticulum. They function in NH₂-terminal proteolytic processing, as shown for the yeast STE24 gene product. This gene is required for the correct processing of α -factor, a yeast pheromone. Family M48 peptidases also appear to be required for some prenylation reactions, mediating COOH-terminal CAAX processing. Prenylation reactions are believed to be involved in the regulation of protein-protein and protein-membrane interactions. As an example, RAS GTPase activity is regulated in part by localization to the inner side of the plasma membrane upon prenylation. In plants, proteins from this family could be involved in pollen-stigma interactions such as those mediating self-pollination vs. outcrossing, or could be members of several secondary metabolism pathways.

Ref: Fujimura-Kamada et al. (1997) J Cell Biol. 136: 271-85. Tam et al. (1998) J Cell Biol. 142: 635-49.

Z. DNA Pol Viral N

The DNA pol Viral N domain is located at the N-terminal region of DNA polymerase isolated from several retroid viruses such as the Cauliflower Mosaic Virus. The domain motif has also been found in numerous other species from humans to cyanobacteria. In these

organisms, this motif seems to be associated with two types of sequences; retrotransposons and mitochondrial genes. In the mitochondrial sequences this domain is potentially involved in the self-splicing conducted by group II introns. Various manipulations of this gene in plants allows control of the numerous retrotransposons endogenous to plant genomes or
 5 allows engineering of mitochondrial function, especially to increase efficiency of energy utilization by cells.

REF: Chapdelaine and Bonen (1991) Cell 65: 465-72

Ferat and Miche (1993) Nature 364: 358-61

10 Wilson et al. (1994) 368: 32-8

Cambareri et al. (1994) 242: 658-65

Gaardner et al. (1981) NAR 9: 2871-2888

Cummings et al. (1990) Curr Genet 17: 375-402

Hattori et al. (1986) Nature 321: 625-8

Aa. Calpain_inhib

This domain is found in calpastatin, an inhibitor protein specific for calpain. Calpain is a non-lysosomal calcium-dependent intracellular protease that appears to be involved in the dynamic changes of the cytoskeleton, especially actin-related structures, during early
 20 *Drosophila* embryogenesis [1]. Calpastatins co-exist in cells with calpains and the subcellular distribution of calpastatin is thought to be important to calpain regulation [2]. In plants calpains and calpastatins could be involved in embryogenesis and non-embryogenic organ reiteration. Mutations occurring in calpain inhibitor repeat domains would produce developmental abnormalities such as abnormal leaf, root or flower development.

Refs

1 Emori Y and Saigo K (1994) J Biol Chem 269: 25137-42.

2 Mellgren RL, Lane RD, Mericle MT (1989) Biochim Biophys Acta 999: 71-77.

Ab. chorismate_bind

30 Chorismate binding domains are present in plant anthranilate synthase (AS) genes. AS genes catalyze the first step in the biosynthesis of tryptophan by converting chorismate and L-glutamine to anthranilate, pyruvate and L-glutamate. Some of these genes are involved in

feedback inhibition by tryptophan [1] while some are feedback insensitive [2]. In Arabidopsis, two AS genes have overlapping, but different distributions. One of these AS genes is induced by wounding and bacterial pathogen infiltration [1]. Mutations in the chorismate binding domain would affect the production of tryptophan and could influence the plant's defense system. AS gene products can be used for *in vitro* synthesis of tryptophan and tryptophan derivatives.

Refs

- 1 Niyogi KK, Fink GR (1992) Plant Cell 4: 721-33.
- 2 Song HS, Brotherton JE, Gonzales RA, Wilholm JM (1998) Plant Physiol 117:533-43.

Ac. late protein_L2

Papillomaviruses are encapsulated double stranded DNA viruses. Plants are susceptible to infection by double stranded DNA viruses such as Cauliflower Mosaic virus (CaMV). The coat proteins in these plant viruses are critical to the virus life cycle within the plant. For example, the coat protein of CaMV is thought to be involved in intra- and inter-cellular movement within the plant [1]. Engineering of proteins having similarity to papillomavirus coat proteins may enable the production of plants having better resistance to natural plant double stranded DNA viruses.

Refs

- 1 Thompson SR, Melcher U (1993) J Gen Virol 74: 1141-8.

Ad. Peptidase_M41

Proteins belonging to this peptidase family are metalloproteases that bind zinc as a cofactor and are integral membrane proteins. They seem to be involved in the degradation of carboxy-terminal-tagged cytoplasmic proteins. In plants, these proteins are located in the thylakoid membranes of the chloroplasts, their expression is light regulated and they are thought to be involved in degradation of soluble stromal proteins and turn-over of thylakoid proteins [1]. Manipulation of expression and structure of these proteins would have effects on the efficiency of photosynthesis and the development of chloroplasts.

Refs

1 Lindahl M, Tabak s, Cseke L, Pichersky E, Andersson B, Adam Z (1996) J Biol Chem 271: 29329-34.

5 Ae. UPF0051

There is some evidence that, in plants, proteins in this family are involved in ATP synthesis in chloroplasts [1, 2]. Mutations in these proteins or altering their expression would affect the efficiency of photosynthesis and energy production.

10 Refs

1 Kostrzewa M, Zetsche K (1992) J Mol Biol 227: 961-70.
2 Kostrzewa M, Zetsche K (1993) Plant Mol Biol 23: 67-76

Af. E7

5 Papillomaviruses are encapsulated double stranded DNA viruses. The Papillomavirus early protein 7 (E7) is known as a potent immortalizing and transforming agent. Transformation by E7 is thought to be mediated by the physical association of E7 with cellular proteins regulating entry into the cell cycle [1]. The result is entry into the cell cycle and suppression of terminal differentiation in mammalian cells. Thus, engineering of proteins having
20 similarity to papillomavirus E7 protein enables the production of plants having altered cellular proliferation characteristics and possibly altered morphology. For example, overexpression of E7-like proteins would be expected to result in proliferation of cells of the tissue in which the E7 protein is expressed, perhaps with suppression of differentiation events. Thus, for example, overexpression of E7-like proteins in meristem cells can result in
25 taller plants and suppression of leafing and/or flowering.

Refs

1 Zwerschke W, Jansen-Durr P Adv Cancer Res 2000;78:1-29

30 Ag. Peptidase U7

This protein is known to be an integral membrane protein in the cyanobacterium Synechocystis where it functions to digest cleaved signal peptides [1]. This activity is necessary to maintain proper secretion of mature proteins across the membrane. In higher

plants this protein may be present in the plastid or chloroplast membranes where it would function by enabling protein movement into and out of the chloroplasts. Mutations in this protein would be expected to affect the development of plastids, including chloroplasts, or alter the energy transfer system within the chloroplasts, thereby affecting growth and development.

Refs

- 1 Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirose M, Sugiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S (1996) DNA Res 3:109-36.

Ah. 5'-3' Exonuclease

The 5'-3' exonuclease domain is one found in bacterial DNA polymerases I and in yeast DNA repair enzymes such as Exonuclease I. Yeast Exo I is involved in mitotic recombination and also includes a domain that interacts with the mismatch repair protein MSH2. The 5'-3' exonuclease domain is also present in XPG DNA repair enzymes in humans and in yeast RAD9 protein. Defects in XPG proteins result in Xeroderma Pigmentosum. Thus defects in 5'-3' exonuclease domain-containing proteins in plants are expected to lead to defects in DNA repair and corresponding high spontaneous and inducible mutation rates. Consensus sequence:

```
IMKKKLLLVDGSSLAFFALPPLTNSAGEPTNAVYGFLKMLIKLIEQEQPTHIAVV
FDAKAKTFRHELVEGYKAGRAP
TPDELREQUIPLIKELLDALGIPLLEVAGYEADDVIGTLAKLAEKEGYEVLIVTGDRDLL
QLVSDHVTVIITKKGIAEFTL
FTPEAVIEKYGLTPEQIIDYKALMGDSSDNIPGVKGIGEKTAACKLLQEYGSLEGYANL
DKLKGKKLREKLLAHKEDAKL
SRDLATIKTDVPLDLTLDDLRLPDPDRDALDLLFDE
```

Ref:

- Fiorentini P. et al. RT. Mol. Cell. Biol. 17:2764-2773(1997).
 Tishkoff et al. Cancer Res. 0:0-0(1998).
 Macinnes M.A. et al. Mol. Cell. Biol. 13:6393-6402(1993).

Table A

Pfam	Prosite	Full Name	Description
3_5_exonuclease		3'-5' exonuclease	<p>Accession number: PF01612</p> <p>Definition: 3'-5' exonuclease</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_659 (release 4.1)</p> <p>Gathering cutoffs: -11 -11</p> <p>Trusted cutoffs: -10.70 -10.70</p> <p>Noise cutoffs: -24.50 -24.50</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 85137890</p> <p>Reference Title: Structure of large fragment of Escherichia coli DNA polymerase I complexed with dTMP.</p> <p>Reference Author: Ollis DL, Brick P, Hamlin R, Xuong NG, Steitz TA;</p> <p>Reference Location: Nature 1985;313:762-766</p> <p>Reference Number: [2]</p> <p>Reference Medline: 98060913</p> <p>Reference Title: The proofreading domain of Escherichia coli DNA polymerase</p> <p>Reference Title: I and other DNA and/or RNA exonuclease domains.</p> <p>Reference Author: Moser MJ, Holley WR, Chatterjee A, Mian IS;</p> <p>Reference Location: Nucleic Acids Res 1997;25:5110-5118</p> <p>Reference Number: [3]</p> <p>Reference Medline: 98361165</p> <p>Reference Title: Replication focus-forming activity 1 and the Werner syndrome gene product</p> <p>Reference Author: Yan H, Chen CY, Kobayashi R, Newport J;</p> <p>Reference Location: Nat Genet 1998;19:375-378</p> <p>Reference Number: [4]</p> <p>Reference Medline: 97434221</p> <p>Reference Title: The Werner syndrome protein is a DNA helicase</p> <p>Reference Author: Gray MD, Shen JC, Kamath-Loeb AS, Blank A, Sopher BL,</p> <p>Reference Author: Martin GM, Oshima J, Loeb LA;</p> <p>Reference Location: Nat Genet 1997;17:100-103.</p> <p>Reference Number: [5]</p> <p>Reference Medline: 97370026</p> <p>Reference Title: DNA helicase activity in Werner's syndrome gene product synthesized in a baculovirus system.</p> <p>Reference Author: Suzuki N, Shimamoto A, Imamura O, Kuromitsu J, Kitao S,</p> <p>Reference Author: Goto M, Furuichi Y;</p> <p>Reference Location: Nucleic Acids Res 1997;25:2973-2978.</p> <p>Database Reference: SCOP, 1dpi; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR002562;</p> <p>Database Reference: PDB; 1kfd A; 348; 518;</p> <p>Database Reference: PDB; 1d8y A; 348; 518;</p> <p>Database Reference: PDB; 1d9d A; 348; 518;</p> <p>Database Reference: PDB; 1d9f A; 348; 518;</p> <p>Database Reference: PDB; 1kfs A; 348; 518;</p> <p>Database Reference: PDB; 1kln A; 348; 518;</p> <p>Database Reference: PDB; 1krp A; 348; 518;</p> <p>Database Reference: PDB; 1ksp A; 348; 518;</p> <p>Database Reference: PDB; 1qsl A; 348; 518;</p> <p>Database Reference: PDB; 2kfn A; 348; 518;</p> <p>Database Reference: PDB; 2kfs A; 348; 518;</p> <p>Database Reference: PDB; 2kzm A; 348; 518;</p> <p>Database Reference: PDB; 2kzz A; 348; 518;</p> <p>Comment: This domain is responsible for the 3'-5' exonuclease proofreading</p> <p>Comment: activity of E. coli DNA polymerase I (polI) and other enzymes,</p> <p>Comment: it catalyses the hydrolysis of unpaired or mismatched nucleotides.</p> <p>Comment: This domain consists of the amino-terminal half of the Klenow fragment</p> <p>Comment: in E. coli polI it is also found in the Werner syndrome helicase</p>

842

			<p>Comment: (WRN), focus forming activity 1 protein (FFA-1) and ribonuclease D</p> <p>Comment: (RNase D).</p> <p>Comment: Werner syndrome is a human genetic disorder causing premature aging;</p> <p>Comment: the WRN protein has helicase activity in the 3'-5' direction [4,5].</p> <p>Comment: The FFA-1 protein is required for formation of a replication foci</p> <p>Comment: and also has helicase activity; it is a homologue of the WRN</p> <p>Comment: protein [3]</p> <p>Comment: RNase D is a 3'-5' exonuclease involved in tRNA processing</p> <p>Comment: Also found in this family is the autoantigen PM/Scl thought to be</p> <p>Comment: involved in polymyositis-scleroderma overlap syndrome.</p> <p>Number of members: 41</p>
3HCDH	PDOC00065	3-hydroxyacyl-CoA dehydrogenase signature	<p>3-hydroxyacyl-CoA dehydrogenase (EC 1.1.1.35) (HCDH) [1] is an enzyme involved in fatty acid metabolism, it catalyzes the reduction of 3-hydroxyacyl-CoA to 3-oxoacyl-CoA. Most eukaryotic cells have 2 fatty-acid beta-oxidation systems, one located in mitochondria and the other in peroxisomes. In peroxisomes 3-hydroxyacyl-CoA dehydrogenase forms, with enoyl-CoA hydratase (ECH) and 3,2-trans-enoyl-CoA isomerase (ECI) a multifunctional enzyme where the N-terminal domain bears the hydratase/isomerase activities and the C-terminal domain the dehydrogenase activity. There are two mitochondrial enzymes: one which is monofunctional and the other which is, like its peroxisomal counterpart, multifunctional.</p> <p>In <i>Escherichia coli</i> (gene <i>fadB</i>) and <i>Pseudomonas fragi</i> (gene <i>faoA</i>) HCDH is part of a multifunctional enzyme which also contains an ECH/ECI domain as well as a 3-hydroxybutyryl-CoA epimerase domain [2]</p> <p>The other proteins structurally related to HCDH are</p> <ul style="list-style-type: none"> - Bacterial 3-hydroxybutyryl-CoA dehydrogenase (EC 1.1.1.157) which reduces 3-hydroxybutanoyl-CoA to acetoacetyl-CoA [3] - Eye lens protein lambda-crystallin [4], which is specific to lagomorphes (such as rabbit). <p>There are two major region of similarities in the sequences of proteins of the HCDH family, the first one located in the N-terminal, corresponds to the NAD-binding site, the second one is located in the center of the sequence. We have chosen to derive a signature pattern from this central region.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [DNE]-x(2)-[GA]-F-[LIVMFY]-x-[NT]-R-x(3)-[PA]-[LIVMFY](2)-x(5)-[LIVMFYCT]-[LIVMFY]-x(2)-[GV]</p> <p>Sequences known to belong to this class detected by the pattern ALL</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update July 1998 / Pattern and text revised</p> <p>References</p> <p>[1] Birktoff J.J., Holden H.M., Hamlin R., Xuong N.-H., Banaszak L.J. Proc. Natl. Acad. Sci. U.S.A. 84:8262-8266(1987).</p> <p>[2] Nakahigashi K., Inokuchi H. Nucleic Acids Res. 18:4937-4937(1990).</p> <p>[3] Mullany P., Clayton C.L., Pallen M.J., Slone R., Al-Saleh A., Tabaqchali S. FEMS Microbiol. Lett. 124:61-67(1994).</p>

			[4] Mulders J.W.M., Hendriks W., Blankesteyn W.M., Bloemendal H., de Jong W W. J. Biol. Chem. 263:15462-15466(1988).
4HPPD_C		4-hydroxyphenylpyruvate dioxygenase C terminal domain	Accession number: PF01626 Definition: 4-hydroxyphenylpyruvate dioxygenase C terminal domain Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1116 (release 4.1) Gathering cutoffs: -35 -35 Trusted cutoffs: -25.80 -25.80 Noise cutoffs: -44.90 -44.90 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 93279307 Reference Title: Human 4-hydroxyphenylpyruvate dioxygenase Primary structure and chromosomal localization of the gene Reference Author: Ruetschi U, Dellsen A, Sahlin P, Stenman G, Rymo L. Reference Author: Lindstedt S: Reference Location: Eur J Biochem 1993;213:1081-1089. Database Reference: INTERPRO: IPR002887, Comment: 4-Hydroxyphenylpyruvic acid dioxygenase (HPD) is an important enzyme Comment: in tyrosine catabolism in most organisms. A genetic deficiency in Comment: this enzyme in humans and mice leads to hereditary tyrosinemia type 3. Comment: The identity of the C-terminus of the HPD makes this part of the Comment: molecule a candidate for a functional role in the catalytic process Comment: [1]. This region is found as a separate protein Swiss:Q49717 that Comment: is somewhat different from HPD and may have a different but related Comment: protein function (Unpublished observation Bateman A) Number of members: 28
5_3_exonuclease		5'-3' exonuclease domain	The 5'-3' exonuclease domain is one found in bacterial DNA polymerases I and in yeast DNA repair enzymes such as Exonuclease I. Yeast Exo I is involved in mitotic recombination and also includes a domain that interacts with the mismatch repair protein MSH2. The 5'-3' exonuclease domain is also present in XPG DNA repair enzymes in humans and in yeast RAD9 protein. Defects in XPG proteins result in Xeroderma Pigmentosum. Thus defects in 5'-3' exonuclease domain-containing proteins in plants are expected to lead to defects in DNA repair and corresponding high spontaneous and inducible mutation rates. Consensus sequence: IMKKKLLLDVGSSSLAFRAFFALPPLTNSAGEPTNAVYGFGLKMLIKLIEQEQPTHIA VVFDAKAKTRHELYEGYKAGRAP TPDELREQIPLIKELLDALGIPLLVAGYEADDVIGTLAKLAEKEGYEVLIVTGDR DLLQLVSDHVTVIITKKGIAEFTL FTPEAVIEKYGLTPEQIIDYKALMGDSSDNIPGVKGIGEKTAAKLLQEYGSLEGIY ANLDKLGKGLREKLLAHKEDAKL SRDLATIKTDVPLDLTLDDLRLPDPDRDALDLLFDE Ref: Fiorentini P. et al. RT Mol. Cell. Biol. 17:2764-2773(1997). Tishkoff et al. Cancer Res. 0:0-0(1998) Macinnes M.A. et al. Mol. Cell. Biol. 13:6393-6402(1993)
60s_ribosomal		60s Acidic ribosomal protein	Accession number: PF00428 Definition: 60s Acidic ribosomal protein Author: Finn RD Alignment method of seed: Clustalw Source of seed members: Pfam-B_151 (release 1.0) Gathering cutoffs: 17.17 Trusted cutoffs: 17.80 17.80 Noise cutoffs: 9.30 9.30 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmbuild --seed 0 HMM

844

			<p>Reference Number: [1] Reference Medline: 96282699 Reference Title: Proteins P1, P2, and P0, components of the eukaryotic ribosome stalk. New structural and functional aspects Reference Author: Remacha M, Jimenez-Diaz A, Santos C, Briones E, Zambrano R, Reference Author: Rodriguez Gabriel MA, Guarinos E, Ballesta JP; Reference Location: Biochem Cell Biol 1995;73:959-968. Database Reference: INTERPRO; IPR001813; Database reference: PFAMB; PB002218; Comment: This family includes archaeobacterial L12, eukaryotic P0, P1 and P2. Number of members: 109</p>
6PF2K	PDOC00158	Phosphoglycerate mutase family phosphohistidine signature	<p>Phosphoglycerate mutase (EC 5.4.2.1) (PGAM) and bisphosphoglycerate mutase (EC 5.4.2.4) (BPGM) are structurally related enzymes which catalyze reactions involving the transfer of phospho groups between the three carbon atoms of phosphoglycerate [1,2]. Both enzymes can catalyze three different reactions, although in different proportions:</p> <ul style="list-style-type: none"> - The isomerization of 2-phosphoglycerate (2-PGA) to 3-phosphoglycerate (3-PGA) with 2,3-diphosphoglycerate (2,3-DPG) as the primer of the reaction. - The synthesis of 2,3-DPG from 1,3-DPG with 3-PGA as a primer - The degradation of 2,3-DPG to 3-PGA (phosphatase EC 3.1.3.13 activity). <p>In mammals, PGAM is a dimeric protein. There are two isoforms of PGAM: the M (muscle) and B (brain) forms. In yeast, PGAM is a tetrameric protein. BPGM is a dimeric protein and is found mainly in erythrocytes where it plays a major role in regulating hemoglobin oxygen affinity as a consequence of controlling 2,3-DPG concentration.</p> <p>The catalytic mechanism of both PGAM and BPGM involves the formation of a phosphohistidine intermediate [3].</p> <p>The bifunctional enzyme 6-phosphofructo-2-kinase / fructose-2,6-bisphosphatase (EC 2.7.1.105 and EC 3.1.3.46) (PF2K) [4] catalyzes both the synthesis and the degradation of fructose-2,6-bisphosphate. PF2K is an important enzyme in the regulation of hepatic carbohydrate metabolism. Like PGAM/BPGM, the fructose-2,6-bisphosphatase reaction involves a phosphohistidine intermediate and the phosphatase domain of PF2K is structurally related to PGAM/BPGM.</p> <p>The bacterial enzyme alpha-ribazole-5'-phosphate phosphatase (gene cobC) which is involved in cobalamin biosynthesis also belongs to this family [5].</p> <p>We built a signature pattern around the phosphohistidine residue.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVM]-x-R-H-G-[EQ]-x(3)-N [H is the phosphohistidine residue] Sequences known to belong to this class detected by the pattern ALL, except for Haemophilus influenzae PGAM. Other sequence(s) detected in SWISS-PROT 2.</p> <p>Note some organisms harbor a form of PGAM independent of 2,3-DPG, this enzyme is not related to the family described above [6] Last update November 1995 / Text revised. References [1] Le Boulch P., Joulin V., Garel M -C., Rosa J., Cohen-Solal M. Biochem Biophys. Res. Commun. 156:874-881(1988). [2] White M.F., Fothergill-Gilmore L.A. FEBS Lett. 229:383-387(1988).</p>

			<p>[3] Rose Z.B. Meth. Enzymol. 87:43-51(1982).</p> <p>[4] Bazan J.F., Fletterick R.J., Pilgis S.J. Proc. Natl. Acad. Sci. U.S.A 86:9642-9646(1989)</p> <p>[5] O'Toole G.A , Trzebiatowski J.R., Escalante-Semerena J C. J. Biol. Chem 269:26503-26511(1994).</p> <p>[6] Grana X., De Lecea L . El-Maghrabi M.R , Urena J M., Caellas C , Carreras J.. Puigdomenech P., Pilgis S.J., Climent F. J. Biol. Chem. 267:12797-12803(1992)</p>
7tm_5		7TM chemorecept or	<p>Accession number: PF01604 Definition: 7TM chemoreceptor Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_942 (release 4 1) Gathering cutoffs: -46 -46 Trusted cutoffs -44.30 -44.30 Noise cutoffs: -47 80 -47 80 HMM build command line. hmmbuild -F HMM SEED HMM build command line. hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98248686 Reference Title: Two large families of chemoreceptor genes in the nematodes Reference Title. Caenorhabditis elegans and Caenorhabditis briggsae reveal Reference Title: extensive gene duplication, diversification, movement, and Reference Title: intron loss. Reference Author: Robertson HM; Reference Location: Genome Res 1998,8:449-463. Database Reference INTERPRO, IPR003003; Comment This large family of proteins are related to 7tm_1. Comment: They are 7 transmembrane receptors. This family does not Comment: include all known members, as there are problems with Comment: overlapping specificity with 7tm_1. Comment: This family is greatly expanded in the nematode worm C. Comment: elegans Number of members: 180</p>
Aa_trans		Transmembra ne amino acid transporter protein	<p>Accession number PF01490 Definition: Transmembrane amino acid transporter protein Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_419 (release 4.0) Gathering cutoffs: 25 25 Trusted cutoffs: 150.80 150 80 Noise cutoffs: 3.60 3.60 HMM build command line: hmmbuild -F HMM SEED HMM build command line hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98007977 Reference Title Identification and characterization of the vesicular GABA transporter Reference Title. transporter Reference Author: McIntire SL, Reimer RJ, Schuske K, Edwards RH, Jorgensen Reference Author: EM; Reference Location: Nature 1997;389:870-876. Database Reference INTERPRO: IPR002422; Database reference: PFAMB; PB020912; Comment: This transmembrane region is found in many amino acid transporters Comment: including UNC-47 and MTR. UNC-47 encodes a vesicular amino butyric acid Comment: (GABA) transporter, (VGAT) UNC-47 is predicted to have</p>

			<p>10 transmembrane Comment: domains Swiss:P34579 [1] MTR is a N system amino acid transporter system Comment: protein involved in methyltryptophan resistance Swiss:P38680 Comment: Other members of this family include proline transporters and amino Comment: acid permeases. Number of members: 50</p>
ABC_tran	PDOC00185	ABC transporters family signature	<p>On the basis of sequence similarities a family of related ATP-binding proteins has been characterized [1 to 5]. These proteins are associated with a variety of distinct biological processes in both prokaryotes and eukaryotes, but a majority of them are involved in active transport of small hydrophilic molecules across the cytoplasmic membrane. All these proteins share a conserved domain of some two hundred amino acid residues, which includes an ATP-binding site. These proteins are collectively known as ABC transporters. Proteins known to belong to this family are listed below (references are only provided for recently determined sequences).</p> <p>In prokaryotes.</p> <ul style="list-style-type: none"> - Active transport systems components: alkylphosphonate uptake(phnC/phnK/ phnL), arabinose (araG); arginine (artP); dipeptide (dciAD;dppD/dppF). ferric enterobactin (fepC); ferrichrome (fhuC); galactoside (mglA) glutamine (glnQ); glycerol-3-phosphate (ugpC), glycine betaine/L-proline (proV); glutamate/aspartate (gltL); histidine (hisP); iron(III) (sfuC), iron(III) dicitrate (fecE). lactose (lacK); leucine/isoleucine/valine (braF/braG;livF/livG); maltose (malK); molybdenum (modC); nickel (nikD/ nikE); oligopeptide (amiE/amiF;oppD/oppF), peptide (sapD/sapF); phosphate (pstB); putrescine (potG), ribose (rbsA); spermidine/putrescine (potA); sulfate (cysA); vitamin B12 (btuD). - Hemolysin/leukotoxin export proteins hlyB, cyaB and lktB. - Colicin V export protein cvaB. - Lactococcal export protein lcnC [6]. - Lantibiotic transport proteins nisT (nisin) and spaT (subtilin) - Extracellular proteases B and C export protein prtD. - Alkaline protease secretion protein aprD - Beta-(1.2)-glucan export proteins chvA and ndvA - Haemophilus influenzae capsule-polysaccharide export protein bexA - Cytochrome c biogenesis proteins ccmA (also known as cycV and helA). - Polysialic acid transport protein kpsT. - Cell division associated ftsE protein (function unknown) - Copper processing protein nosF from Pseudomonas stutzeri. - Nodulation protein nodI from Rhizobium (function unknown) - Escherichia coli proteins cydC and cydD. - Subunit A of the ABC excision nuclease (gene uvrA). - Erythromycin resistance protein from Staphylococcus epidermidis (gene msrA). - Tylosin resistance protein from Streptomyces fradiae (gene tirC) [7] - Heterocyst differentiation protein (gene hetA) from Anabaena PCC 7120 - Protein P29 from Mycoplasma hyorhinis, a probable component of a high affinity transport system. - yhbG, a putative protein whose gene is linked with ntrA in many bacteria such as Escherichia coli, Klebsiella pneumoniae, Pseudomonas putida, Rhizobium meliloti and Thiobacillus ferrooxidans. - Escherichia coli and related bacteria hypothetical proteins yabJ, yadG, yagC, ybbA, ycjW, yddA, yehX, yejF, yheS, yhiG, yhiH, yjcW, yjjK, yojl, yrbF and ytfR. <p>In eukaryotes:</p> <ul style="list-style-type: none"> - The multidrug transporters (Mdr) (P-glycoprotein), a family of closely related proteins which extrude a wide variety of drugs out of the cell (for a review see [8]). - Cystic fibrosis transmembrane conductance regulator (CFTR), which is most probably involved in the transport of chloride ions. - Antigen peptide transporters 1 (TAP1, PSF1, RING4, HAM-1, mtp1) and 2 (TAP2, PSF2, RING11, HAM-2, mtp2), which are involved in the transport of antigens from the cytoplasm to a membrane-bound compartment for association with MHC class I molecules.

			<p>- 70 Kd peroxisomal membrane protein (PMP70)</p> <p>- ALDP, a peroxisomal protein involved in X-linked adrenoleukodystrophy [9].</p> <p>- Sulfonyleurea receptor [10], a putative subunit of the B-cell ATP-sensitive potassium channel</p> <p>- Drosophila proteins white (w) and brown (bw), which are involved in the import of ommatidium screening pigments.</p> <p>- Fungal elongation factor 3 (EF-3).</p> <p>- Yeast STE6 which is responsible for the export of the a-factor pheromone.</p> <p>- Yeast mitochondrial transporter ATM1</p> <p>- Yeast MDL1 and MDL2.</p> <p>- Yeast SNQ2.</p> <p>- Yeast sporidesmin resistance protein (gene PDR5 or STS1 or YDR1).</p> <p>- Fission yeast heavy metal tolerance protein hmt1. This protein is probably involved in the transport of metal-bound phytochelatins.</p> <p>- Fission yeast brefeldin A resistance protein (gene bfr1 or hba2).</p> <p>- Fission yeast leptomycin B resistance protein (gene pmd1).</p> <p>- mbpX, a hypothetical chloroplast protein from Liverwort.</p> <p>- Prestalk-specific protein tagB from slime mold. This protein consists of two domains: a N-terminal subtilase catalytic domain (see <PDOC00125>) and a C-terminal ABC transporter domain.</p> <p>As a signature pattern for this class of proteins, we use a conserved region which is located between the 'A' and the 'B' motifs of the ATP-binding site.</p> <p style="text-align: center;">Consensus pattern [LIVMFYC]-[SA]-[SAPGLVFYKQH]-G-[DENQMW]-[KRQASPCLIMFW]-[KRNQSTAVM]-[KRACLVM]-[LIVMFYPAN]-{PHY}-[LIVMFW]-[SAGCLVPI]-{FYWHP}-[KRHP]-[LIVMFYWSTA] Sequences known to belong to this class detected by the pattern ALL, except for 25 sequences. Other sequence(s) detected in SWISS-PROT 42. Note the ATP-binding region is duplicated in araG, mdl, msrA, rbsA, tlrC, uvrA, yefF, Mdr's, CFTR, pmd1 and in EF-3. In some of those proteins, the above pattern only detect one of the two copies of the domain. Note the proteins belonging to this family also contain one or two copies of the ATP-binding motifs 'A' and 'B' (see <PDOC00017>). July 1998 / Text revised.</p> <p>[1] Higgins C.F., Hyde S.C., Mimmack M.M., Gileadi U., Gill D.R., Gallagher M.P. J. Bioenerg. Biomembr. 22:571-592(1990)</p> <p>[2] Higgins C.F., Gallagher M.P., Mimmack M.M., Pearce S.R. BioEssays 8:111-116(1988).</p> <p>[3] Higgins C.F., Hiles I.D., Salmond G.P.C., Gill D.R., Downie J.A., Evans I.J., Holland I.B., Gray L., Buckels S.D., Bell A.W., Hermodson M.A. Nature 323:448-450(1986)</p> <p>[4] Doolittle R.F., Johnson M.S., Husain I., van Houten B., Thomas D.C., Sancar A. Nature 323:451-453(1986).</p> <p>[5] Blight M.A., Holland I.B. Mol. Microbiol. 4:873-880(1990)</p> <p>[6] Stoddard G.W., Petzel J.P., van Belkum M.J., Kok J., McKay L.L. Appl. Environ. Microbiol. 58:1952-1961(1992)</p> <p>[7] Rosteck P.R. Jr., Reynolds P.A., Hershberger C.L. Gene 102:27-32(1991)</p> <p>[8] Gottesman M.M., Pastan I. J. Biol. Chem. 263:12163-12166(1988).</p> <p>[9] Valle D., Gaertner J. Nature 361:682-683(1993).</p> <p>[10] Aguilar-Bryan L., Nichols C.G., Wechsler S.W., Clement J.P. IV, Boyd A.E. III, Gonzalez G., Herrera-Sosa H., Nguy K., Bryan J., Nelson D.A. Science 268:423-426(1995).</p>
ABC2_memb rane	PDOC00692	ABC-2 type transport system integral membrane proteins	<p>Integral membrane components of a number of bacterial active transport systems have been shown to be evolutionary related and to form a distinct family [1,2]. These proteins are:</p> <p>- Escherichia coli kpsM, involved in polysialic acid export.</p>

		signature	<ul style="list-style-type: none"> - Haemophilus influenzae bexB, involved in polyribosylribitol phosphate capsule polysaccharide export. - Salmonella typhi vexB, involved in translocation of the Vi polysaccharide. - Neisseria meningitidis ctrC, involved in polyneuraminic acid capsule polysaccharide export - Rhizobiaceae nodulation protein J (gene nodJ), probably involved in exporting a modified beta-1,4-linked N-acetylglucosamine oligosaccharide - Streptomyces peucetius drrB, involved in exporting the antibiotics daunorubicin and doxorubicin. - Klebsiella pneumoniae O-antigen exprt system protein rfbA. - Yersinia enterocolitica O-antigen exprt system protein rfbD. - Escherichia coli hypothetical protein yadH. - Escherichia coli hypothetical protein yhhJ. <p>The molecular size of these proteins is around 30 Kd. They are thought to contain six transmembrane regions. They either form homooligomeric channels or associate with another type of transmembrane protein to form heteroligomers. Transport systems in which they participate are energized by an ATP-binding protein that belongs to the ABC transporter family. The designation 'ABC-2' has been proposed [1] for these transport systems</p> <p>As a signature pattern, we selected a conserved region located in the C-terminal section of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIMST]-x(2)-[LIMW]-x(2)-[LIMCA]-[GSTC]-x-[GSAIV]-x(6)-[LIMGA]-[PGSNQ]-x(9,12)-P-[LIMFT]-x-[HRSY]-x(5)-[RQ] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT 2. Last update November 1997 / Pattern and text revised. References [1] Reizer J , Reizer A . Saier M H. Jr. Protein Sci 1.1326-1332(1992). [2] Vazquez M., Santana O., Quinto C. Mol Microbiol. 8:369-377(1993).</p>
ABC-3		ABC 3 transport family	<p>Members of this family include receptors that mediate transmembrane signalling. These receptors can bind to a number of factors including amphiregulin, epidermal growth factor, gp30, heparin-binding egf, insulin, insulin-like growth factor I and II, neuregulins, transforming growth factor-alpha and, and vaccinia virus growth</p> <p>Signal transduction is mediated by catalytic activity of tyrosine kinase, such as ATP + A protein tyrosine = ADP + protein tyrosine phosphate Typically, such signal transduction have been implicated in metabolic and developmental changes, including cell fate and differentiation. Examples include instruction of follicle cells to follow a dorsal pathway of development rather than the default ventral pathway. may also bind the spitz protein. References describing these family members and their biological activities.</p> <p>Abbot et al., J. Biol. Chem. 267:10759-10763(1992), Araki et al. J. Biol. Chem 262 16186-16191(1987); Aroian et al., EMBO J. 13:360-366(1994); Aroian et al., Nature 348:693-699(1990); Barbetti et al., Diabetes 41:408-415(1992); Bargmann et al., Nature 319:226-230(1986); Cama et al., J. Biol. Chem. 268:8060-8069(1993); Cama et al., J. Clin. Endocrinol. Metab. 73:894-901(1991); Carrera et al., Hum. Mol. Genet. 2:1437-1441(1993); Clifford et al., Genetics 137:531-550(1994); Cocozza et al., Diabetes 41:521-526(1992), Cooke et al., Biochem. Biophys. Res. Commun. 177:1113-1120(1991); Coussens et al., Science 230:1132-1139(1985); Dickens et al., Biochem. Biophys. Res. Commun. 186:244-250(1992); Ebina et al., Cell 40:747-758(1985); Ebina et al., Proc. Natl. Acad. Sci. U.S.A. 84:704-708(1987); Ehsani et al., Genomics 15:426-429(1993); Elbein et al., Diabetes 42:429-434(1993); Elbein, Diabetes 38:737-743(1989); Fujita-Yamaguchi et al., Protein Seq. Data Anal. 1:3-6(1987); Gullick et al., EMBO J. 11:43-48(1992), Haruta et al., Diabetes 42:1837-1844(1993), Hubbard et al., EMBO J. 16 5572-5581(1997).</p>

			<p>Hubbard et al., Nature 372:746-754(1994); Iwanishi et al., Diabetologia 36:414-422(1993); Kadowaki et al., J. Clin. Invest. 86:254-264(1990); Kadowaki et al., Science 240 787-790(1988); Kim et al., Diabetologia 35:261-266(1992); Klinkhamer et al., EMBO J. 8:2503-2507(1989); Kusari et al., J. Biol. Chem. 266:5260-5267(1991); Lai et al., Neuron 6:691-704(1991); Lax et al., Mol. Cell. Biol. 8:1970-1978(1988); Lebrun et al., J. Biol. Chem. 268:11272-11277(1993); Lee et al., Oncogene 8:3403-3410(1993); Lesokhin et al., Dev. Biol. 205:129-144(1999); Livneh et al., Cell 40:599-607(1985); Longo et al., Proc. Natl Acad. Sci. U S.A. 90:60-64(1993); McKeon et al., Mol. Endocrinol. 4:647-656(1990); Moller et al., J. Biol. Chem. 265:14979-14985(1990); Moller et al., Mol. Endocrinol. 4:1183-1191(1990); Odawara et al., Science 245:66-68(1989); Raz et al., Genetics 129:191-201(1991); Sakai et al., J. Mol. Biol. 256:548-555(1996); Schaeffer et al., Biochem. Biophys. Res. Commun. 189:650-653(1992); Schejter et al., Cell 46:1091-1101(1986); Seino et al., Biochem. Biophys. Res. Commun. 159:312-316(1989); Seino et al., Diabetes 39:123-128(1990); Semba et al., Proc. Natl Acad. Sci. U.S.A. 82:6497-6501(1985); Shier et al., J. Biol. Chem. 264:14605-14608(1989); Taira et al., Science 245 63-66(1989); Tewari et al., J. Biol. Chem. 264 16238-16245(1989); Ullrich et al., Nature 313:756-761(1985); Ullrich et al., EMBO J. 5:2503-2512(1986); van der Vorm et al., Diabetologia 36:172-174(1993); van der Vorm et al., J. Biol. Chem. 267:66-71(1992); Wadsworth et al., Nature 314:178-180(1985); White et al., Cell 54:641-649(1988); Xu et al., J. Biol. Chem. 265:18673-18681(1990); Yamamoto et al., Nature 319:230-234(1986), and Yoshimasa et al., Science 240:784-787(1988)</p>
ACAT		Sterol O-acyltransferase	<p>Accession number PF01800 Definition: Sterol O-acyltransferase Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1454 (release 4.2) Gathering cutoffs: 25 25 Trusted cutoffs: 112.80 112.80 Noise cutoffs: -128.10 -128.10 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98434592 Reference Title: Characterization of two human genes encoding acyl coenzyme Reference Title: A:cholesterol acyltransferase-related enzymes. Reference Author: Oelkers P, Behari A, Cromley D, Billheimer JT, Sturley SL; Reference Location: J Biol Chem 1998;273:26765-26771. Reference Number: [2] Reference Medline: 98434590 Reference Title: Identification of a form of acyl-CoA:cholesterol Reference Title: acyltransferase specific to liver and intestine in nonhuman Reference Title: primates. Reference Author: Anderson RA, Joyce C, Davis M, Reagan JW, Clark M, Shelness Reference Author: GS, Rudel LL; Reference Location: J Biol Chem 1998;273:26747-26754. Reference Number: [3] Reference Medline: 96243137 Reference Title: Sterol esterification in yeast: a two-gene process. Reference Author: Yang H, Bard M, Bruner DA, Gleeson A, Deckelbaum RJ. Reference Author: Aljinovic G, Pohl TM, Rothstein R, Sturley SL, Reference Location: Science 1996;272 1353-1356. Database Reference: INTERPRO; IPR002688, Comment: Sterol O-acyltransferases or acyl-coa:cholesterol acyltransferase Comment: (ACAT) EC.2.3.1.26 is a transmembrane protein that catalyses the Comment: esterification of cholesterol to its cholesterol ester storage Comment: form. Number of members: 21</p>
ACPS		4'-phosphopantetheinyl transferase superfamily	<p>Accession number PF01648 Definition: 4'-phosphopantetheinyl transferase superfamily Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1679 (release 4.1)</p>

850

			<p>Gathering cutoffs: 0.0</p> <p>Trusted cutoffs: 0.60 0.60</p> <p>Noise cutoffs: -4.00 -4.00</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96027548</p> <p>Reference Title: Cloning, overproduction, and characterization of the</p> <p>Reference Title: Escherichia coli holo-acyl carrier protein synthase.</p> <p>Reference Author: Lambalot RH, Walsh CT;</p> <p>Reference Location: J Biol Chem 1995;270:24658-24661.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97144264</p> <p>Reference Title: A new enzyme superfamily - the phosphopantetheinyl</p> <p>Reference Title: transferases.</p> <p>Reference Author: Lambalot RH, Gehring AM, Flugel RS, Zuber P. LaCelle M,</p> <p>Reference Author: Marahiel MA, Reid R, Khosla C, Walsh CT.</p> <p>Reference Location: Chem Biol 1996;3:923-936.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 10581256</p> <p>Reference Title: Crystal structure of the surfactin synthetase-activating</p> <p>Reference Title: enzyme sfp: a prototype of the 4'-phosphopantetheinyl</p> <p>Reference Title: transferase superfamily [In Process Citation]</p> <p>Reference Author: Reuter K, Mofid MR, Marahiel MA, Ficner R;</p> <p>Reference Location: EMBO J 1999;18:6823-6831.</p> <p>Database Reference: INTERPRO; IPR002582,</p> <p>Database reference: PFAMB; PB007908;</p> <p>Database reference: PFAMB; PB041384;</p> <p>Comment: Members of this family transfers the</p> <p>Comment: 4'-phosphopantetheine (4'-PP) moiety from coenzyme A</p> <p>Comment: (CoA) to</p> <p>Comment: the invariant serine of pp-binding This post-translational</p> <p>Comment: modification renders holo-ACP capable of acyl group</p> <p>Comment: activation</p> <p>Comment: via thioesterification of the cysteamine thiol of 4'-PP [1].</p> <p>Comment: This superfamily consists of two subtypes: The ACPS type</p> <p>Comment: such as Swiss:P24224 and the Sfp type such as</p> <p>Comment: Swiss:P39135.</p> <p>Comment: The structure of the Sfp type is known [3], which shows</p> <p>Comment: the</p> <p>Comment: active site accommodates a magnesium ion The most</p> <p>Comment: highly</p> <p>Comment: conserved regions of the alignment are involved in binding</p> <p>Comment: the magnesium ion.</p> <p>Number of members: 46</p>
ACT		ACT domain	<p>Accession number: PF01842</p> <p>Definition: ACT domain</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Manual</p> <p>Source of seed members: Bateman A</p> <p>Gathering cutoffs: 25.0</p> <p>Trusted cutoffs: 26.10 0.50</p> <p>Noise cutoffs: 24.50 24.50</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 95236205</p> <p>Reference Title: The allosteric ligand site in the Vmax-type cooperative</p> <p>Reference Title: enzyme phosphoglycerate dehydrogenase.</p> <p>Reference Author: Schuller DJ, Grant GA, Banaszak LJ;</p> <p>Reference Location: Nat Struct Biol 1995;2:69-76.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 99241053</p> <p>Reference Title: Gleaning non-trivial structural, functional and</p> <p>Reference Title: evolutionary information about proteins by iterative</p> <p>Reference Title: database searches.</p> <p>Reference Author: Aravind L, Koonin EV;</p> <p>Reference Location: J Mol Biol 1999;287:1023-1040.</p> <p>Database Reference: SCOP; 1psd, fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR002912;</p> <p>Database Reference: PDB; 1phz A; 35; 110;</p> <p>Database Reference: PDB; 2phm A; 35; 110;</p>

851

			<p>Database Reference PDB, 1psd A; 338; 410;</p> <p>Database Reference PDB; 1psd B; 338; 410;</p> <p>Database reference: PFAMB; PB001977;</p> <p>Database reference: PFAMB; PB008097;</p> <p>Database reference: PFAMB; PB010480;</p> <p>Database reference: PFAMB; PB011031;</p> <p>Database reference: PFAMB; PB031880;</p> <p>Database reference: PFAMB; PB038464;</p> <p>Database reference: PFAMB; PB040963;</p> <p>Database reference: PFAMB; PB041518;</p> <p>Database reference: PFAMB; PB041667;</p> <p>Comment: This family of domains generally have a regulatory role.</p> <p>Comment: ACT domains are linked to a wide range of metabolic</p> <p>Comment: enzymes that are regulated by amino acid concentration.</p> <p>Comment: Pairs of ACT domains bind specifically to a particular</p> <p>Comment: amino acid leading to regulation of the linked enzyme.</p> <p>Comment: The ACT domain is found in:</p> <p>Comment: D-3-phosphoglycerate dehydrogenase EC.1.1.1 95</p> <p>Swiss.P08328,</p> <p>Comment: which is inhibited by serine [1]</p> <p>Comment: Aspartokinase EC:2.7.2.4 Swiss.P53553, which is</p> <p>regulated by lysine.</p> <p>Comment: Acetolactate synthase small regulatory subunit</p> <p>Swiss.P00894,</p> <p>Comment: which is inhibited by valine.</p> <p>Comment: Phenylalanine-4-hydroxylase EC.1.14.16.1 Swiss:P00439,</p> <p>which</p> <p>Comment: is regulated by phenylalanine</p> <p>Comment: Prephenate dehydrogenase EC:4.2.1.51 Swiss.P21203.</p> <p>Comment: formyltetrahydrofolate deformylase EC.3.5.1.10,</p> <p>Swiss.P37051,</p> <p>Comment: which is activated by methionine and inhibited by glycine</p> <p>Comment: GTP pyrophosphokinase EC:2.7.6.5 Swiss:P11585.</p> <p>Number of members: 177</p>
Acyl-ACP_TE		Acyl-ACP thioesterase	<p>Accession number PF01643</p> <p>Definition: Acyl-ACP thioesterase</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_928 (release 4.1)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 91.70 91 70</p> <p>Noise cutoffs: -192.80 -192.80</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96068671</p> <p>Reference Title: Modification of the substrate specificity of an acyl-acyl</p> <p>Reference Title: carrier protein thioesterase by protein engineering.</p> <p>Reference Author: Yuan L, Voelker TA, Hawkins DJ,</p> <p>Reference Location: Proc Natl Acad Sci U S A 1995;92:10639-10643.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 92320297</p> <p>Reference Title: Fatty acid biosynthesis redirected to medium chains in</p> <p>Reference Title: transgenic oilseed plants.</p> <p>Reference Author: Voelker TA, Worrell AC, Anderson L, Bleibaum J, Fan</p> <p>C,</p> <p>Reference Author: Hawkins DJ, Radke SE, Davies HM;</p> <p>Reference Location: Science 1992;257:72-74.</p> <p>Database Reference INTERPRO; IPR002864,</p> <p>Comment: This family consists of various acyl-acyl carrier protein</p> <p>(ACP)</p> <p>Comment: thioesterases (TE) these terminate fatty acyl group</p> <p>extension via</p> <p>Comment: hydrolyzing an acyl group on a fatty acid [1].</p> <p>Number of members: 30</p>
Acyltransferase		Acyltransferase	<p>Accession number: PF01553</p> <p>Definition: Acyltransferase</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_128 (release 4.0)</p> <p>Gathering cutoffs: 8 8</p> <p>Trusted cutoffs: 14 40 14 40</p>

852

			<p>Noise cutoffs: 2 50 2.50 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97411131 Reference Title: Barth syndrome may be due to an acyltransferase deficiency. Reference Author: Neuwald AF; Reference Location: Curr Biol 1997;7:465-466. Reference Number: [2] Reference Medline: 96224398 Reference Title: A novel X-linked gene, G4.5. is responsible for Barth syndrome. Reference Author: Bione S, D'Adamo P, Maestrini E, Gedeon AK. Bolhuis PA, Reference Author: Toniolo D; Reference Location: Nat Genet 1996;12:385-389. Database Reference: INTERPRO, IPR002123, Database reference: PFAMB; PB009622; Database reference: PFAMB; PB009717; Database reference: PFAMB; PB033259; Database reference: PFAMB; PB041102; Database reference: PFAMB; PB041638; Comment: This family contains acyltransferases involved in phospholipid Comment: biosynthesis and other proteins of unknown function [1] This Comment: family also includes tafazzin Swiss:Q16635, the Barth syndrome Comment: gene [2]. Number of members: 74</p>
Adaptin_N		Adaptin N terminal region	<p>Accession number: PF01602 Definition: Adaptin N terminal region Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_491 (release 4.0) Gathering cutoffs: 12 12 Trusted cutoffs: 15 50 15 50 Noise cutoffs: 9.00 9.00 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97409270 Reference Title: Linking cargo to vesicle formation: receptor tail Reference Title: interactions with coat proteins Reference Author: Kirchhausen T, Bonifacino JS, Riezman H; Reference Location: Curr Opin Cell Biol 1997;9:488-495. Reference Number: [2] Reference Medline: 89202379 Reference Title: Structural and functional division into two domains of the large (100- to 115-kDa) chains of the clathrin-associated protein complex AP-2. Reference Title: RAKirchhausen T, Nathanson KL, Matsui W, Varsberg A, Chow Reference Author: EP, Burne C, Keen JH, Davis AE; Reference Location: Proc Natl Acad Sci U S A 1989;86:2612-2616. Database Reference: INTERPRO, IPR002553, Database reference: PFAMB; PB040953; Comment: This family consists of the N terminal region of various alpha, Comment: beta and gamma subunits of the AP-1, AP-2 and AP-3 adaptor Comment: protein complexes. The adaptor protein (AP) complexes are involved in Comment: the formation of clathrin-coated pits and vesicles [1]. Comment: The N-terminal region of the various adaptor proteins (APs) is constant Comment: by comparison to the C-terminal which is variable within members of the Comment: AP-2 family[2]; and it has been proposed that this constant region Comment: interacts with another uniform component of the coated vesicles [2].</p>

853

			Number of members: 66
ALAD	PDOC00153	Delta-aminolevulinic acid dehydratase active site	<p>Delta-aminolevulinic acid dehydratase (EC 4.2.1.24) (ALAD) [1] catalyzes the second step in the biosynthesis of heme, the condensation of two molecules of 5-aminolevulinate to form porphobilinogen. The enzyme is an oligomer composed of eight identical subunits. Each of the subunits binds an atom of zinc or of magnesium (in plants). A lysine has been implicated in the catalytic mechanism [2]. The sequence of the region in the vicinity of the active site residue is conserved in ALAD from various prokaryotic and eukaryotic species</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-x-D-x-[LIVM](2)-[IV]-K-P-[GSA]-x(2)-Y [K is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1995 / Pattern and text revised.</p> <p>References [1] Li J -M., Russell C.S., Cosloy S D. Gene 75:177-184(1989).</p> <p>[2] Gibbs P N.B., Jordan P.M. Biochem. J. 236:447-451(1986).</p>
Aldolase	PDOC00144	KDPG and KHG aldolases active site signatures	<p>4-hydroxy-2-oxoglutarate aldolase (EC 4.1.3.16) (KHG-aldolase) catalyzes the interconversion of 4-hydroxy-2-oxoglutarate into pyruvate and glyoxylate. Phospho-2-dehydro-3-deoxygluconate aldolase (EC 4.1.2.14) (KDPG-aldolase) catalyzes the interconversion of 6-phospho-2-dehydro-3-deoxy-D-gluconate into pyruvate and glyceraldehyde 3-phosphate.</p> <p>These two enzymes are structurally and functionally related [1]. They are both homotrimeric proteins of approximately 220 amino-acid residues. They are class I aldolases whose catalytic mechanism involves the formation of a Schiff-base intermediate between the substrate and the epsilon-amino group of a lysine residue. In both enzymes, an arginine is required for catalytic activity.</p> <p>We developed two signature patterns for these enzymes. The first one contains the active site arginine and the second, the lysine involved in the Schiff-base formation.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-[LIVM]-x(3)-E-[LIV]-T-[LF]-R [R is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for <i>Bacillus subtilis</i> KDPG-aldolase which has Thr instead of Arg in the active site.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern G-x(3)-[LIVMF]-K-[LF]-F-P-[SA]-x(3)-G [K is involved in Schiff-base formation]</p> <p>Sequences known to belong to this class detected by the pattern ALL</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Patterns and text revised.</p> <p>References [1] Vlahos C J., Dekker E E. J. Biol. Chem. 263 11683-11691(1988).</p>
Alpha_L_fucosidases	PDOC00324	Alpha-L-fucosidase	<p>Alpha-L-fucosidase (EC 3.2.1.51) [1] is a lysosomal enzyme responsible for hydrolyzing the alpha-1,6-linked fucose joined to the reducing-end N-acetylglucosamine of the carbohydrate moieties of glycoproteins. Deficiency of alpha-L-fucosidase results in the lysosomal storage disease fucosidosis</p>

			<p>A cysteine residue is important for the activity of the enzyme. There is only one cysteine conserved between the sequence of mammalian alpha-L-fucosidase and that of the slime mold Dictyostelium discoideum. We have derived a pattern from the region around that conserved cysteine.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern P-x(2)-L-x(3)-K-W-E-x-C [C is the putative active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note these proteins belong to family 29 in the classification of glycosyl hydrolases [2,E1].</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References [1] Fisher K J , Aronson N.N. Jr Biochem. J. 264:695-701(1989).</p> <p>[2] Hennissat B. Biochem. J. 280:309-316(1991).</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?glycosid.txt</p>
Amino_oxidase	Flavin containing amine oxidase		<p>Accession number: PF01593</p> <p>Definition: Flavin containing amine oxidase</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_606 (release 4 1)</p> <p>Gathering cutoffs: -110 -110</p> <p>Trusted cutoffs: -110.00 -110.00</p> <p>Noise cutoffs: -111.80 -111.80</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98258926</p> <p>Reference Title: Maize polyamine oxidase: primary structure from protein and cDNA sequencing.</p> <p>Reference Author: Tavladoraki P, Schinina ME, Cecconi F, Agostino SD, Manera</p> <p>Reference Author: F, Rea G, Mariottini P, Federico R, Angelini R.</p> <p>Reference Location: FEBS Lett 1998;426:62-66.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97306298</p> <p>Reference Title: A key amino acid responsible for substrate selectivity of monoamine oxidase A and B.</p> <p>Reference Author: Tsugeno Y, Ito A,</p> <p>Reference Location: J Biol Chem 1997;272:14033-14036.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 95287865</p> <p>Reference Title: Cloning, sequencing and heterologous expression of the monoamine oxidase gene from Aspergillus niger.</p> <p>Reference Author: Schilling B, Lerch K;</p> <p>Reference Location: Mol Gen Genet 1995;247:430-438.</p> <p>Database Reference: SCOP, 1b37; fa: [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO: IPR002937;</p> <p>Database Reference: PDB; 1b37 A; 14; 455;</p> <p>Database Reference: PDB; 1b5q A; 14; 455;</p> <p>Database Reference: PDB; 1b37 B; 14; 455;</p> <p>Database Reference: PDB; 1b37 C; 14; 455;</p> <p>Database Reference: PDB; 1b5q B; 14; 455;</p> <p>Database Reference: PDB; 1b5q C; 14; 455;</p> <p>Database reference: PFAMB; PB017518;</p> <p>Database reference: PFAMB; PB024839;</p> <p>Database reference: PFAMB; PB040747;</p> <p>Comment: This family consists of various amine oxidases, including</p>

			<p>maze polyamine oxidase (PAO) [1] and various flavin containing monoamine oxidases (MAO). The aligned region includes the flavin binding site of these enzymes.</p> <p>In vertebrates MAO plays an important role regulating the intracellular levels of amines via their oxidation; these include various neurotransmitters, neurotoxins and trace amines [2]. In lower eukaryotes such as aspergillus and in bacteria the main role of amine oxidases is to provide a source of ammonium [3].</p> <p>PAOs in plants, bacteria and protozoa oxidase spermidine and spermine to an aminobutylal, diaminopropane and hydrogen peroxide and are involved in the catabolism of polyamines [1]</p> <p>Other members of this family include tryptophan 2-monooxygenase, putrescine oxidase, corticosteroid binding proteins and antibacterial glycoproteins.</p> <p>Number of members: 58</p>
ANF_receptor	PDOC00430	Natriuretic peptides receptors signature	<p>Natriuretic peptides are hormones involved in the regulation of fluid and electrolyte homeostasis. These hormones stimulate the intracellular production of cyclic GMP as a second messenger.</p> <p>Currently, three types of natriuretic peptide receptors are known [1,2]. Two express guanylate cyclase activity. GC-A (or ANP-A) which seems specific to atrial natriuretic peptide (ANP), and GC-B (or ANP-B) which seems to be stimulated more effectively by brain natriuretic peptide (BNP) than by ANP. The third receptor (ANP-C) is probably responsible for the clearance of ANP from the circulation and does not play a role in signal transduction</p> <p>GC-A and GC-B are plasma membrane-bound proteins that share the following topology: an N-terminal extracellular domain which acts as the ligand binding region, then a transmembrane domain followed by a large cytoplasmic C-terminal region that can be subdivided into two domains. a protein kinase-like domain (see <PDOC00100>) that appears important for proper signalling and a guanylate cyclase catalytic domain (see <PDOC00425>). The topology of ANP-C is different: like GC-A and -B it possesses an extracellular ligand-binding region and a transmembrane domain, but its cytoplasmic domain is very short</p> <p>We developed a pattern from the ligand-binding region of natriuretic peptide receptors based on a highly conserved region located in the N-terminal part of the domain</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-P-x-C-x-Y-x-A-A-x-V-x-R-x(3)-H-W Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE Last update May 1991 / First entry. References [1] Garbers D.L. New Biol. 2:499-504(1990). [2] Schulz S., Chinkers M., Garbers D.L. FASEB J 2:2026-2035(1989).</p>
Apocytochrome_F	PDOC00169	Cytochrome c family heme-binding site	<p>In proteins belonging to cytochrome c family [1], the heme group is covalently attached by thioether bonds to two conserved cysteine residues. The consensus</p>

856

		signature	<p>sequence for this site is Cys-X-X-Cys-His and the histidine residue is one of the two axial ligands of the heme iron. This arrangement is shared by all proteins known to belong to cytochrome c family, which presently includes cytochromes c. c', c1 to c6. c550 to c556, cc3/Hmc, cytochrome f and reaction center cytochrome c</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-{CPWHF}-{CPWR}-C-H-{CFYW}</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for four cytochrome c's which lack the first thioether bond.</p> <p>Other sequence(s) detected in SWISS-PROT 454.</p> <p>Note. some cytochrome c's have more than a single bound heme group c4 has 2, c7 has 3. c3 has 4, the reaction center has 4, and cc3/Hmc has 16 !</p> <p>Last update June 1992 / Text revised.</p> <p>References [1] Mathews F.S. Prog. Biophys. Mol. Biol. 45:1-56(1985).</p>
arf	PDOC00781 PDOC00017 PDOC01020	ADP- ribosylation factors family signature, ATP/GTP- binding site motif A (P- loop); ATP phosphoribos- yltransferase signature PROSITE cross- reference(s)	<p>ADP-ribosylation factors (ARF) [1,2,3,4] are 20 Kd GTP-binding proteins involved in protein trafficking. They may modulate vesicle budding and uncoating within the Golgi apparatus. ARF's also act as allosteric activators of cholera toxin ADP-ribosyltransferase activity. They are evolutionary conserved and present in all eukaryotes. At least six forms of ARF are present in mammals and three in budding yeast. The ARF family also includes proteins highly related to ARF's but which lack the cholera toxin cofactor activity, they are collectively known as ARL's (ARF-like).</p> <p>ARD1 is a 64 Kd mammalian protein of unknown biological function that contains an ARF domain at its C-terminal extremity.</p> <p>Proteins from the ARF family are generally included in the RAS 'superfamily' of small GTP-binding proteins [5], but they are only slightly related to the other RAS proteins. They also differ from RAS proteins in that they lack cysteine residues at their C-termini and are therefore not subject to prenylation. The ARFs are N-terminally myristoylated (the ARLs have not yet been shown to be modified in such a fashion).</p> <p>As a signature pattern, we selected a conserved region in the C-terminal part of ARF's and ARL's.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [HRQT]-x-[FYWI]-x-[LIVM]-x(4)-A-x(2)-G-x(2)-[LIVM]-x(2)-[GSA]-[LIVMF]-x-[WK]-[LIVM]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for 4 sequences.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note proteins belonging to this family also contain a copy of the ATP/GTP-binding motif 'A' (P-loop) (see <PDOC00017>).</p> <p>Expert(s) to contact by email Kahn R.A. rkahn@bimcore.emory.edu</p> <p>Last update November 1997 / Pattern and text revised Cell Signal. 4:367-399(1993). References [1] Boman A.L., Kahn R.A. Trends Biochem. Sci. 20:147-150(1995).</p> <p>[2] Moss J., Vaughan M.</p> <p>[3]</p>

		<p>Moss J., Vaughan M. Prog. Nucleic Acid Res. Mol. Biol. 45:47-65(1993)</p> <p>[4] Amor J.C., Harrison D.H., Kahn R A., Ringe D. Nature 372:704-708(1994)</p> <p>[5] Valencia A., Chardin P., Wittinghofer A., Sander C Biochemistry 30:4637-4648(1991)</p> <p>From sequence comparisons and crystallographic data analysis it has been shown [1,2,3,4,5,6] that an appreciable proportion of proteins that bind ATP or GTP share a number of more or less conserved sequence motifs. The best conserved of these motifs is a glycine-rich region, which typically forms a flexible loop between a beta-strand and an alpha-helix. This loop interacts with one of the phosphate groups of the nucleotide. This sequence motif is generally referred to as the 'A' consensus sequence [1] or the 'P-loop' [5].</p> <p>There are numerous ATP- or GTP-binding proteins in which the P-loop is found. We list below a number of protein families for which the relevance of the presence of such motif has been noted:</p> <ul style="list-style-type: none"> - ATP synthase alpha and beta subunits (see <PDOC00137>). - Myosin heavy chains. - Kinesin heavy chains and kinesin-like proteins (see <PDOC00343>) - Dynamins and dynamin-like proteins (see <PDOC00362>) - Guanylate kinase (see <PDOC00670>). - Thymidine kinase (see <PDOC00524>). - Thymidylate kinase (see <PDOC01034>) - Shikimate kinase (see <PDOC00868>). - Nitrogenase iron protein family (nifH/frxC) (see <PDOC00580>). - ATP-binding proteins involved in 'active transport' (ABC transporters) [7] (see <PDOC00185>). - DNA and RNA helicases [8,9,10]. - GTP-binding elongation factors (EF-Tu, EF-1alpha, EF-G, EF-2, etc) - Ras family of GTP-binding proteins (Ras, Rho, Rab, Ral, Ypt1, SEC4, etc). - Nuclear protein ran (see <PDOC00859>). - ADP-ribosylation factors family (see <PDOC00781>). - Bacterial dnaA protein (see <PDOC00771>). - Bacterial recA protein (see <PDOC00131>). - Bacterial recF protein (see <PDOC00539>). - Guanine nucleotide-binding proteins alpha subunits (Gi, Gs, Gt, G0, etc). - DNA mismatch repair proteins mutS family (See <PDOC00388>). - Bacterial type II secretion system protein E (see <PDOC00567>). <p>Not all ATP- or GTP-binding proteins are picked-up by this motif. A number of proteins escape detection because the structure of their ATP-binding site is completely different from that of the P-loop. Examples of such proteins are the E1-E2 ATPases or the glycolytic kinases. In other ATP- or GTP-binding proteins the flexible loop exists in a slightly different form; this is the case for tubulins or protein kinases. A special mention must be reserved for adenylate kinase, in which there is a single deviation from the P-loop pattern: in the last position Gly is found instead of Ser or Thr.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [AG]-x(4)-G-K-[ST] Sequences known to belong to this class detected by the pattern a majority. Other sequence(s) detected in SWISS-PROT in addition to the proteins listed above, the 'A' motif is also found in a number of other proteins. Most of these proteins probably bind a nucleotide, but others are definitively not ATP- or GTP-binding (as for example chymotrypsin, or human ferritin light chain). Expert(s) to contact by email Koonin E.V. koonin@ncbi.nlm.nih.gov</p> <p>Last update July 1999 / Text revised. References [1]</p>
--	--	---

858

		<p>Walker J.E., Saraste M., Runswick M.J., Gay N.J. EMBO J. 1:945-951(1982).</p> <p>[2] Moller W , Amons R FEBS Lett. 186 1-7(1985).</p> <p>[3] Fry D C., Kuby S.A., Mildvan A.S Proc. Natl. Acad. Sci. U.S.A 83:907-911(1986).</p> <p>[4] Dever T.E., Glynnias M.J., Merrick W C. Proc. Natl. Acad. Sci. U S.A 84:1814-1818(1987)</p> <p>[5] Saraste M., Sibbald P R., Wittinghofer A. Trends Biochem. Sci. 15:430-434(1990)</p> <p>[6] Koonin E.V. J. Mol. Biol 229:1165-1174(1993).</p> <p>[7] Higgins C.F , Hyde S.C , Mimmack M.M , Gileadi U. Gill D R., Gallagher M.P J Bioenerg. Biomembr 22:571-592(1990)</p> <p>[8] Hodgman T C. Nature 333:22-23(1988) and Nature 333:578-578(1988) (Errata).</p> <p>[9] Linder P , Lasko P., Ashburner M., Leroy P., Nielsen P.J. Nishi K., Schnier J , Slonimski P.P. Nature 337:121-122(1989)</p> <p>[10] Gorbalenya A.E , Koonin E.V., Donchenko A.P , Blinov V M Nucleic Acids Res. 17 4713-4730(1989).</p> <p>ATP phosphoribosyltransferase (EC 2.4.2.17) is the enzyme that catalyzes the first step in the biosynthesis of histidine in bacteria, fungi and plants. It is a protein of about 23 to 32 Kd. As a signature pattern we selected a region located in the C-terminal part of this enzyme.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern E-x(5)-G-x-[SAG]-x(2)-[IV]-x-D-[LIV]-x(2)-[ST]-G-x-T-[LM] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE. Last update July 1998 / First entry.</p>
ArgJ		<p>ArgJ family</p> <p>Accession number: PF01960 Definition: ArgJ family Author: Enright A. Ouzounis C, Bateman A Alignment method of seed. Clustalw Source of seed members. Enright A Gathering cutoffs: 25 25 Trusted cutoffs: 258.70 99.60 Noise cutoffs: 7.10 7.10 HMM build command line. hmmbuild -f HMM SEED HMM build command line. hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 93232760 Reference Title: Primary structure, partial purification and regulation of key enzymes of the acetyl cycle of arginine biosynthesis in Reference Title: Bacillus stearothermophilus: dual function of ornithine acetyltransferase. Reference Author: Sakanyan V, Charlier D, Legrain C, Kochikyan A, Mett I, Reference Author: Pierard A, Glansdorff N; Reference Location: J Gen Microbiol 1993;139:393-402. Database Reference INTERPRO; IPR002813,</p>

859

			<p>Comment: Members of the ArgJ family catalyse the first EC:2.3.1.35 and</p> <p>Comment: fifth steps EC:2.3.1.1 in arginine biosynthesis.</p> <p>Number of members: 22</p>
Armadillo_seg		Armadillo/beta-catenin-like repeats	<p>Accession number: PF00514</p> <p>Definition: Armadillo/beta-catenin-like repeats</p> <p>Author: Bateman A, Chris Ponting, Joerg Schultz, Peer Bork</p> <p>Alignment method of seed: Manual</p> <p>Source of seed members: SMART</p> <p>Gathering cutoffs: 24 0</p> <p>Trusted cutoffs: 24.10 0.00</p> <p>Noise cutoffs: 20 70 20.20</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97442350</p> <p>Reference Title: Three-dimensional structure of the armadillo repeat region of beta-catenin.</p> <p>Reference Author: Huber AH, Nelson WJ, Weis WI;</p> <p>Reference Location: Cell 1997;90:871-882.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 96107551</p> <p>Reference Title: Signal transduction of beta-catenin</p> <p>Reference Author: Gumbiner BM;</p> <p>Reference Location: Curr Opin Cell Biol 1995;7:634-640.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 97454713</p> <p>Reference Title: Armadillo and dTCF: a marriage made in the nucleus</p> <p>Reference Author: Cavallo R, Rubenstein D, Peifer M;</p> <p>Reference Location: Curr Opin Genet Dev 1997;7:459-466</p> <p>Reference Number: [4]</p> <p>Reference Medline: 94082295</p> <p>Reference Title: Association of the APC tumor suppressor protein with catenins.</p> <p>Reference Author: Su LK, Vogelstein B, Kinzler KW</p> <p>Reference Location: Science 1993;262:1734-1737.</p> <p>Reference Number: [5]</p> <p>Reference Medline: 94082294</p> <p>Reference Title: Association of the APC gene product with beta-catenin.</p> <p>Reference Author: Rubinfeld B, Souza B, Albert I, Muller O, Chamberlain SH,</p> <p>Reference Author: Masiarz FR, Munemitsu S, Polakis P,</p> <p>Reference Location: Science 1993;262:1731-1734.</p> <p>Reference Number: [6]</p> <p>Reference Medline: 91084846</p> <p>Reference Title: The segment polarity gene armadillo encodes a functionally</p> <p>Reference Title: modular protein that is the Drosophila homolog of human plakoglobin.</p> <p>Reference Author: Peifer M, Wieschaus E,</p> <p>Reference Location: Cell 1990;63 1167-1176.</p> <p>Database Reference: SCOP; 3bct; fa: [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: EXPERT, Chris.Ponting@human-anatomy.oxford.ac.uk;</p> <p>Database reference: SMART; ARM,</p> <p>Database Reference: INTERPRO; IPR000225,</p> <p>Database Reference: PDB; 1ee5 A; 417; 457;</p> <p>Database Reference: PDB; 1bk5 A; 417; 457;</p> <p>Database Reference: PDB; 1bk5 B; 417; 457;</p> <p>Database Reference: PDB; 1bk6 A; 417; 457;</p> <p>Database Reference: PDB; 1bk6 B; 417; 457;</p> <p>Database Reference: PDB; 1ee4 A; 417; 457;</p> <p>Database Reference: PDB; 1ee4 B; 417; 457;</p> <p>Database Reference: PDB; 1ejl I; 409; 449;</p> <p>Database Reference: PDB; 1ejy I; 409; 449;</p> <p>Database Reference: PDB; 1ial A; 409; 449;</p> <p>Database Reference: PDB; 1ee5 A; 246; 286;</p> <p>Database Reference: PDB; 1bk5 A; 246; 286;</p> <p>Database Reference: PDB; 1bk5 B; 246; 286;</p> <p>Database Reference: PDB; 1bk6 A; 246; 286;</p> <p>Database Reference: PDB; 1bk6 B; 246; 286;</p> <p>Database Reference: PDB; 1ee4 A; 246; 286;</p> <p>Database Reference: PDB; 1ee4 B; 246; 286;</p>

	Database Reference	PDB; 1ejl I; 241; 280;
	Database Reference	PDB; 1ejy I; 241; 280;
	Database Reference	PDB; 1ial A; 241; 280;
	Database Reference	PDB; 1ee5 A; 288; 328;
	Database Reference	PDB; 1bk5 A; 288; 328;
	Database Reference	PDB; 1bk5 B; 288; 328;
	Database Reference	PDB; 1bk6 A; 288; 328;
	Database Reference	PDB; 1bk6 B; 288; 328;
	Database Reference	PDB; 1ee4 A; 288; 328;
	Database Reference	PDB; 1ee4 B; 288; 328;
	Database Reference	PDB; 1ejl I; 282; 322;
	Database Reference	PDB; 1ejy I; 282; 322;
	Database Reference	PDB; 1ial A; 282; 322;
	Database Reference	PDB; 1ejl I; 151; 191;
	Database Reference	PDB; 1ejy I; 151; 191;
	Database Reference	PDB; 1ial A; 151; 191;
	Database Reference	PDB; 1ee5 A; 162; 202;
	Database Reference	PDB; 1bk5 A; 162; 202;
	Database Reference	PDB; 1bk5 B; 162; 202;
	Database Reference	PDB; 1bk6 A; 162; 202;
	Database Reference	PDB; 1bk6 B; 162; 202;
	Database Reference	PDB; 1ee4 A; 162; 202;
	Database Reference	PDB; 1ee4 B; 162; 202;
	Database Reference	PDB; 1ee5 A; 330; 370;
	Database Reference	PDB; 1bk5 A; 330; 370;
	Database Reference	PDB; 1bk5 B; 330; 370;
	Database Reference	PDB; 1bk6 A; 330; 370;
	Database Reference	PDB; 1bk6 B; 330; 370;
	Database Reference	PDB; 1ee4 A; 330; 370;
	Database Reference	PDB; 1ee4 B; 330; 370;
	Database Reference	PDB; 1ejl I; 324; 364;
	Database Reference	PDB; 1ejy I; 324; 364;
	Database Reference	PDB; 1ial A; 324; 364;
	Database Reference	PDB; 1ee5 A; 372; 412;
	Database Reference	PDB; 1bk5 A; 372; 412;
	Database Reference	PDB; 1bk5 B; 372; 412;
	Database Reference	PDB; 1bk6 A; 372; 412;
	Database Reference	PDB; 1bk6 B; 372; 412;
	Database Reference	PDB; 1ee4 A; 372; 412;
	Database Reference	PDB; 1ee4 B; 372; 412;
	Database Reference	PDB; 1ejl I; 366; 406;
	Database Reference	PDB; 1ejy I; 366; 406;
	Database Reference	PDB; 1ial A; 366; 406;
	Database Reference	PDB; 1ejl I; 108; 149;
	Database Reference	PDB; 1ejy I; 108; 149;
	Database Reference	PDB; 1ial A; 108; 149;
	Database Reference	PDB; 1ee5 A; 119; 160;
	Database Reference	PDB; 1bk5 A; 119; 160;
	Database Reference	PDB; 1bk5 B; 119; 160;
	Database Reference	PDB; 1bk6 A; 119; 160;
	Database Reference	PDB; 1bk6 B; 119; 160;
	Database Reference	PDB; 1ee4 A; 119; 160;
	Database Reference	PDB; 1ee4 B; 119; 160;
	Database Reference	PDB; 3bct ; 583; 623;
	Database Reference	PDB; 2bct ; 583; 623;
	Database Reference	PDB; 3bct , 391, 429;
	Database Reference	PDB; 2bct ; 391, 429;
	Database Reference	PDB; 3bct ; 224; 264;
	Database Reference	PDB; 2bct , 224; 264;
	Database Reference	PDB; 3bct , 431; 473;
	Database Reference	PDB; 2bct ; 431; 473;
	Database Reference	PDB; 3bct , 350; 390;
	Database Reference	PDB; 2bct , 350; 390;
	Database Reference	PDB; 1ejl I; 193; 238;
	Database Reference	PDB; 1ejy I; 193; 238;
	Database Reference	PDB; 1ial A; 193; 238;
	Database Reference	PDB; 1ee5 A; 204; 244;
	Database Reference	PDB; 1bk5 A; 204; 244;
	Database Reference	PDB; 1bk5 B; 204; 244;
	Database Reference	PDB; 1bk6 A; 204; 244;
	Database Reference	PDB; 1bk6 B; 204; 244;
	Database Reference	PDB; 1ee4 A; 204; 244;
	Database Reference	PDB; 1ee4 B; 204; 244;
	Database Reference	PDB; 1ibr D; 399; 437;

861

			<p>Database Reference PDB; 1ibr B; 399; 437; Database Reference PDB; 1qgk A; 399; 437; Database Reference PDB; 1qgr A; 399; 437; Database reference: PFAMB; PB002221; Database reference: PFAMB; PB002617; Database reference: PFAMB; PB004638; Database reference: PFAMB; PB012310; Database reference: PFAMB; PB040528; Database reference: PFAMB; PB041028; Comment: Approx 40 amino acid repeat. Tandem repeats form super-helix of helices Comment: that is proposed to mediate interaction of beta-catenin with its ligands Comment: CAUTION: This family does not contain all known armadillo repeats. Number of members: 597</p>
ATP synt_B_c	PDOC00137	ATP synthase alpha and beta subunits signature	<p>ATP synthase (proton-translocating ATPase) (EC 3.6.1.34) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria, and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), and a catalytic core, called coupling factor CF(1). The former acts as a proton channel; the latter is composed of five subunits, alpha, beta, gamma, delta and epsilon. The sequences of subunits alpha and beta are related and both contain a nucleotide-binding site for ATP and ADP. The beta chain has catalytic activity, while the alpha chain is a regulatory subunit</p> <p>Vacuolar ATPases [3] (V-ATPases) are responsible for acidifying a variety of intracellular compartments in eukaryotic cells. Like F-ATPases, they are oligomeric complexes of a transmembrane and a catalytic sector. The sequence of the largest subunit of the catalytic sector (70 Kd) is related to that of F-ATPase beta subunit, while a 60 Kd subunit, from the same sector, is related to the F-ATPases alpha subunit [4].</p> <p>Archaeobacterial membrane-associated ATPases are composed of three subunits. The alpha chain is related to F-ATPases beta chain and the beta chain is related to F-ATPases alpha chain [4]</p> <p>A protein highly similar to F-ATPase beta subunits is found [5] in some bacterial apparatus involved in a specialized protein export pathway that proceeds without signal peptide cleavage. This protein is known as flil in <i>Bacillus</i> and <i>Salmonella</i>, Spa47 (mxlB) in <i>Shigella flexneri</i>, HrpB6 in <i>Xanthomonas campestris</i> and yscN in <i>Yersinia</i> virulence plasmids.</p> <p>In order to detect these ATPase subunits, we took a segment of ten amino-acid residues, containing two conserved serines, as a signature pattern. The first serine seems to be important for catalysis - in the ATPase alpha chain at least - as its mutagenesis causes catalytic impairment.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern P-[SAP]-[LIV]-[DNH]-x(3)-S-x-S [The first S is a putative active site residue] Sequences known to belong to this class detected by the pattern ALL, except for the archaeobacterium <i>Sulfolobus acidocaldarius</i> ATPase alpha chain where the first Ser is replaced by Gly. Other sequence(s) detected in SWISS-PROT 37</p> <p>Note F-ATPase alpha and beta subunits, V-ATPase 70 Kd subunit and the archaeobacterial ATPase alpha subunit also contain a copy of the ATP-binding motifs A and B (see <PDOC00017>)</p> <p>Last update November 1997 / Pattern and text revised</p> <p>References [1] Futai M., Noumi T., Maeda M Annu. Rev. Biochem. 58.111-136(1989).</p>

			<p>[2] Senior A.E Physiol. Rev. 68:177-231(1988).</p> <p>[3] Nelson N. J. Bioenerg. Biomembr 21:553-571(1989).</p> <p>[4] Gogarten J P , Kibak H., Dittich P , Taiz L., Bowman E.J., Bowman B.J., Manolson M F , Poole R.J., Date T., Oshima T., Konishi J., Denda K., Yoshida M Proc. Natl Acad. Sci. U.S.A 86:6661-665(1989).</p> <p>[5] Dreyfus G., Williams A W., Kawagishi I., MacNab R.M. J. Bacteriol. 175.3131-3138(1993).</p>
ATP-synt_D		ATP synthase subunit D	<p>Accession number. PF01813 Definition: ATP synthase subunit D Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members Pfam-B_1304 (release 4.2) Gathering cutoffs: 25 25 Trusted cutoffs: 157.80 157.80 Noise cutoffs: -79.90 -79.90 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 96324968 Reference Title. Subunit structure and organization of the genes of the A1A0 Reference Title: ATPase from the Archaeon Methanosarcina mazei Go1 Reference Author. Wilms R, Freiberg C, Wegerle E, Meier I, Mayer F, Muller V; Reference Location: J Biol Chem 1996;271:18843-18852. Reference Number: [2] Reference Medline: 95132627 Reference Title. A bovine cDNA and a yeast gene (VMA8) encoding the subunit Reference Title. D of the vacuolar H(+)-ATPase. Reference Author: Nelson H, Mandrian S, Nelson N; Reference Location: Proc Natl Acad Sci U S A 1995;92:497-501. Database Reference INTERPRO; IPR002699, Comment: This is a family of subunit D form various ATP synthases Comment: including V-type H+ transporting and Na+ dependent. Comment: Subunit D is suggested to be an integral part of the catalytic sector of the V-ATPase [2] Number of members: 21</p>
B56		Protein phosphatase 2A regulatory B subunit (B56 family)	<p>Accession number. PF01603 Definition: Protein phosphatase 2A regulatory B subunit (B56 family) Author: Bateman A Alignment method of seed: Clustalw Source of seed members Pfam-B_984 (release 4 1) Gathering cutoffs: 11 11 Trusted cutoffs: 17.80 17.80 Noise cutoffs: 5.50 5.50 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 96064678 Reference Title. Identification of a new family of protein phosphatase 2A regulatory subunits. Reference Author: McCright B, Virshup DM; Reference Location: J Biol Chem 1995;270:26123-26128. Database Reference INTERPRO; IPR002554; Comment: Protein phosphatase 2A (PP2A) is a major intracellular protein Comment: phosphatase that regulates multiple aspects of cell growth and metabolism Comment: The ability of this widely distributed heterotrimeric enzyme to act on a</p>

863

			<p>Comment: diverse array of substrates is largely controlled by the nature of its</p> <p>Comment: regulatory B subunit. There are multiple families of B subunits (See also</p> <p>Comment: PR55), this family is called the B56 family [1].</p> <p>Number of members 34</p>
Bac_export_1		Bacterial export proteins, family 1	<p>Accession number. PF01311</p> <p>Definition: Bacterial export proteins, family 1</p> <p>Author: Finn RD, Bateman A</p> <p>Alignment method of seed. Clustalw</p> <p>Source of seed members Pfam-B_1442 (release 3.0)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 37 20 37 20</p> <p>Noise cutoffs: -95.00 -95 00</p> <p>HMM build command line. hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 95113771</p> <p>Reference Title: Caulobacter FliQ and FliR membrane proteins, required for</p> <p>Reference Title: flagellar biogenesis and cell division, belong to a family of virulence factor export proteins.</p> <p>Reference Author Zhuang WY, Shapiro L;</p> <p>Reference Location J Bacteriol 1995.177:343-356.</p> <p>Database Reference INTERPRO, IPR002010.</p> <p>Comment: This family includes the following members,</p> <p>Comment: FliR, MopE, SsaT, YopT, Hrp, HrcT and SpaR</p> <p>Comment: All of these members export proteins, that do not possess signal</p> <p>Comment: peptides. through the membrane Although the proteins that these</p> <p>Comment: exporters move may be different. the exporters are thought to</p> <p>Comment: function in similar ways [1].</p> <p>Number of members: 29</p>
Band 41	PDOC00566	Band 4.1 family domain signatures and profile	<p>A number of cytoskeletal-associated proteins that associate with various proteins at the interface between the plasma membrane and the cytoskeleton contain a conserved N-terminal domain of about 150 amino-acid residues [1.2, 3] The proteins in which such a domain is known to exist are listed below.</p> <ul style="list-style-type: none"> - Band 4 1, which links the spectrin-actin cytoskeleton of erythrocytes to the plasma membrane. Band 4 1 binds with a high affinity to glycophorin and with lower affinity to band 3 protein. - Ezrin (cytovillin or p81), a component of the undercoat of the microvilli plasma membrane. - Moesin, which is probably involved in binding major cytoskeletal structures to the plasma membrane. - Radixin, which seems to play a crucial role in the binding of the barbed end of actin filaments to the plasma membrane in the undercoat of the cell-to-cell adherens junction (AJ). - Talin, which binds with high affinity to vinculin and with low affinity to integrins. Talin is a high molecular weight (270 Kd) cytoskeletal protein concentrated in regions of cell-substratum contact and, in lymphocytes, of cell-cell contacts. - Filopodin, a slime mold protein that binds actin ans which is involved in the control of cell motility and chemotaxis. - Merlin (or schwannomin) Defects in this protein are the cause of type 2 neurofibromatosis (NF2), a predisposition to tumors of the nervous system - Protein NBL4. - Protein-tyrosine phosphatases PTPN3 (PTP-H1) and PTPN4 (PTP-MEG1). <p>Structurally these two very similar enzymes are composed of a N-terminal band 4.1-like domain followed by a central segment of unknown function and a C-terminal catalytic domain (see <PDOC00323>) They could act at junctions between the membrane and the cytoskeleton.</p> <ul style="list-style-type: none"> - Protein-tyrosine phosphatases PTPN14 (PEZ or PTP36) and PTP-D1. PTP-RL10 and PTP2E. These phosphatases also consist of a N-terminal band 4.1-like domain and a C-terminal catalytic domain. The central domain seems to contain a SH3-binding domain - Caenorhabditis elegans protein phosphatase ptp-1.

			<p>Ezrin, moesin, and radixin are highly related proteins, but the other proteins in which this domain is found do not share any region of similarity outside of the domain. In band 4.1 this domain is known to be important for the interaction with glycophorin, an integral membrane protein.</p> <p>We have developed two signature patterns for this domain. one is based on the conserved positions found at the N-terminal extremity of the domain, the second is located in the C-terminal section.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern W-[LIV]-x(3)-[KRQ]-x-[LIVM]-x(2)-[QH]-x(0,2)-[LIVMF]-x(6,8)-[LIVMF]-x(3,5)-F-[FY]-x(2)-[DENS] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [HYW]-x(9)-[DENQSTV]-[SA]-x(3)-[FY]-[LIVM]-x(2)-[ACV]-x(2)-[LM]-x(2)-[FY]-G-x-[DENQST]-[LIVMFYS] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE</p> <p>Sequences known to belong to this class detected by the profile ALL Other sequence(s) detected in SWISS-PROT 7.</p> <p>Note this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so. Expert(s) to contact by email Rees J. jrees@vax.oxford.ac.uk</p> <p>Last update November 1997 / Patterns and text revised; profile added</p> <p>References [1] Rees D.J.G., Ades S.A., Singer S.J., Hynes R.O. Nature 347 685-689(1990).</p> <p>[2] Funayama N., Nagafuchi A., Sato N., Tsukita S., Tsukita S. J. Cell Biol 115:1039-1048(1991)</p> <p>[3] Takeuchi K., Kawashima A., Nagafuchi A., Tsukita S. J. Cell Sci. 107:1921-1928(1994).</p>
biotin_lipoyl	PDOC00167; PDOC00168	Biotin-requiring enzymes; 2-oxo acid dehydrogenases acyltransferase component lipoyl binding	<p>Biotin, which plays a catalytic role in some carboxyl transfer reactions, is covalently attached, via an amide bond, to a lysine residue in enzymes requiring this coenzyme [1,2,3,4]. Such enzymes are:</p> <ul style="list-style-type: none"> - Pyruvate carboxylase (EC 6.4.1.1) - Acetyl-CoA carboxylase (EC 6.4.1.2). - Propionyl-CoA carboxylase (EC 6.4.1.3). - Methylcrotonoyl-CoA carboxylase (EC 6.4.1.4). - Geranoyl-CoA carboxylase (EC 6.4.1.5). - Urea carboxylase (EC 6.3.4.6). - Oxaloacetate decarboxylase (EC 4.1.1.3) - Methylmalonyl-CoA decarboxylase (EC 4.1.1.41) - Glutaconyl-CoA decarboxylase (EC 4.1.1.70). - Methylmalonyl-CoA carboxyl-transferase (EC 2.1.3.1) (transcarboxylase). <p>Sequence data reveal that the region around the biocytin (biotin-lysine) residue is well conserved and can be used as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [GN]-[DEQTR]-x-[LIVMFY]-x(2)-[LIVM]-x-[AIV]-M-K-[LMAT]-x(3)-[LIVM]-x-[SAV] [K is the biotin attachment site] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE.</p>

			<p>Note the domain around the biotin-binding lysine residue is evolutionary related to that around the lipoyl-binding lysine residue of 2-oxo acid dehydrogenase acyltransferases (see <PDOC00168>)</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References</p> <p>[1] Knowles J.R. Annu. Rev. Biochem. 58:195-221(1989).</p> <p>[2] Samols D., Thronton C.G., Murtif V.L., Kumar G.K., Haase F.C., Wood H.G. J. Biol. Chem. 263:6461-6464(1988).</p> <p>[3] Goss N.H., Wood H.G. Meth. Enzymol. 107:261-278(1984).</p> <p>[4] Shenoy B.C., Xie Y., Park V.L., Kumar G.K., Beegen H., Wood H.G., Samols D. J. Biol. Chem. 267:18407-18412(1992).</p> <p>The 2-oxo acid dehydrogenase multienzyme complexes [1,2] from bacterial and eukaryotic sources catalyze the oxidative decarboxylation of 2-oxo acids to the corresponding acyl-CoA. The three members of this family of multienzyme complexes are:</p> <ul style="list-style-type: none"> - Pyruvate dehydrogenase complex (PDC). - 2-oxoglutarate dehydrogenase complex (OGDC). - Branched-chain 2-oxo acid dehydrogenase complex (BCOADC). <p>These three complexes share a common architecture: they are composed of multiple copies of three component enzymes - E1, E2 and E3. E1 is a thiamine pyrophosphate-dependent 2-oxo acid dehydrogenase, E2 a dihydrolipamide acyltransferase, and E3 an FAD-containing dihydrolipamide dehydrogenase.</p> <p>E2 acyltransferases have an essential cofactor, lipoic acid, which is covalently bound via an amide linkage to a lysine group. The E2 components of OGDC and BCOADC bind a single lipoyl group, while those of PDC bind either one (in yeast and in <i>Bacillus</i>), two (in mammals), or three (in <i>Azotobacter</i> and in <i>Escherichia coli</i>) lipoyl groups [3].</p> <p>In addition to the E2 components of the three enzymatic complexes described above, a lipoic acid cofactor is also found in the following proteins:</p> <ul style="list-style-type: none"> - H-protein of the glycine cleavage system (GCS) [4]. GCS is a multienzyme complex of four protein components, which catalyzes the degradation of glycine. H protein shuttles the methylamine group of glycine from the P protein to the T protein. H-protein from either prokaryotes or eukaryotes binds a single lipoic group. - Mammalian and yeast pyruvate dehydrogenase complexes differ from that of other sources, in that they contain, in small amounts, a protein of unknown function - designated protein X or component X. Its sequence is closely related to that of E2 subunits and seems to bind a lipoic group [5]. - Fast migrating protein (FMP) (gene <i>acoC</i>) from <i>Alcaligenes eutrophus</i> [6]. This protein is most probably a dihydrolipamide acyltransferase involved in acetoin metabolism. <p>We developed a signature pattern which allows the detection of the lipoyl-binding site.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [GN]-x(2)-[LIVF]-x(5)-[LIVFC]-x(2)-[LIVFA]-x(3)-K-[STAIV]-[STAVQDN]-x(2)-[LIVMFS]-x(5)-[GCN]-x-[LIVMFY] [K is the lipoyl-binding site] Sequences known to belong to this class detected by the pattern ALL.</p>
--	--	--	---

866

			<p>Other sequence(s) detected in SWISS-PROT 2</p> <p>Note the domain around the lipoyl-binding lysine residue is evolutionary related to that around the biotin-binding lysine residue of biotin requiring enzymes (see <PDOC00167>).</p> <p>Last update November 1995 / Text revised.</p> <p>References</p> <p>[1] Yeaman S.J. Biochem. J. 257:625-632(1989).</p> <p>[2] Yeaman S.J. Trends Biochem. Sci. 11:293-296(1986).</p> <p>[3] Russel G C , Guest J.R Biochim. Biophys. Acta 1076:225-232(1991).</p> <p>[4] Fujiwara K., Okamura-Ikeda K., Motokawa Y. J. Biol. Chem 261:8836-8841(1986).</p> <p>[5] Behal R.H., Browning K.S., Hall T.B., Reed L.J. Proc. Natl Acad. Sci. U.S.A 86:8732-8736(1989).</p> <p>[6] Priefert H., Hein S , Krueger N , Zeh K , Schmidt B. Steinbuechel A. J. Bacteriol. 173 4056-4071(1991).</p>
Biotin synth		Biotin synthase	<p>Accession number PF01792</p> <p>Definition: Biotin synthase</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1407 (release 4.2)</p> <p>Gathering cutoffs: -180 -180</p> <p>Trusted cutoffs -176.30 -176.30</p> <p>Noise cutoffs: -183.90 -183.90</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96312354</p> <p>Reference Title. Cloning, sequencing, and characterization of the Bacillus subtilis biotin biosynthetic operon.</p> <p>Reference Title: subtilis biotin biosynthetic operon.</p> <p>Reference Author: Bower S. Perkins JB, Yocum RR, Howitt CL, Rahaim P, Pero J;</p> <p>Reference Location: J Bacteriol 1996.178:4122-4130.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97074643</p> <p>Reference Title. Two new members of the bio B superfamily cloning, sequencing and expression of bio B genes of</p> <p>Reference Title: sequencing and expression of bio B genes of</p> <p>Methylobacillus</p> <p>Reference Title: flagellatum and Corynebacterium glutamicum.</p> <p>Reference Author: Serebriiskii IG, Vassin VM, Tsygankov YD;</p> <p>Reference Location: Gene 1996,175:15-22.</p> <p>Database Reference INTERPRO; IPR002684,</p> <p>Database reference: PFAMB; PB023954;</p> <p>Database reference: PFAMB; PB040740;</p> <p>Database reference: PFAMB; PB041208;</p> <p>Comment: Biotin synthase EC:2.8.1.6 works with flavodoxin, S-adenosylmethionine,</p> <p>Comment: and possibly cysteine to convert dethiobiotin to biotin [1].</p> <p>Comment: Biotin (vitamin H) is a prosthetic group in enzymes</p> <p>Comment: catalysing</p> <p>Comment: carboxylation and transcarboxylation reactions [2]</p> <p>Number of members: 29</p>
BolA		BolA-like protein	<p>Accession number: PF01722</p> <p>Definition: BolA-like protein</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1996 (release 4.1)</p>

867

			<p>Gathering cutoffs: 23 23 Trusted cutoffs: 23.70 23.70 Noise cutoffs: -16.00 -16.00 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 99291046 Reference Title: The stationary-phase morphogene bolA from Escherichia coli Reference Title: is induced by stress during early stages of growth. Reference Author: Santos JM, Freire P, Vicente M, Arraiano CM; Reference Location: Mol Microbiol 1999;32:789-798. Reference Number: [2] Reference Medline: 90059998 Reference Title: Induction of a growth-phase-dependent promoter triggers transcription of bolA, an Escherichia coli morphogene. Reference Author: Aldea M, Garrido T, Hernandez-Chico C. Vicente M, Kushner Reference Author: SR; Reference Location: EMBO J 1989;8:3923-3931. Database Reference: INTERPRO, IPR002634, Comment: This family consist of the morpho-protein BolA from Comment: E. coli and its various homologs In E. coli over expression of Comment: this protein causes round morphology and may be Comment: involved in Comment: switching the cell between elongation and septation Comment: systems during Comment: cell division [1]. The expression of BolA is growth rate Comment: regulated Comment: and is induced during the transition into the the stationary Comment: phase [1]. BolA is also induced by stress during early Comment: stages of Comment: growth [1] and may have a general role in stress Comment: response. Comment: It has also been suggested that BolA can induce the Comment: transcription Comment: of penicillin binding proteins 6 and 5 [2,1] Number of members: 18</p>
casein_kappa			<p>Accession number PF00997 Definition: Kappa casein Author: Bateman A Alignment method of seed: Clustalw Source of seed members Pfam-B_1298 (release 3.0) Gathering cutoffs: -32 -32 Trusted cutoffs 16 40 16 40 Noise cutoffs: -73.00 -73.00 HMM build command line hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98072500 Reference Title: Nucleotide sequence evolution at the kappa-casein locus: Reference Title: evidence for positive selection within the family Bovidae. Reference Author: Ward TJ, Honeycutt RL, Derr JN; Reference Location: Genetics 1997;147 1863-1872. Database Reference: INTERPRO: IPR000117; Comment: Kappa-casein is a mammalian milk protein involved in a Comment: number of important physiological processes. In the gut, Comment: the ingested protein is split into an insoluble peptide Comment: (para kappa-casein) and a soluble hydrophilic Comment: glycopeptide Comment: (caseinomacropptide). Caseinomacropptide is Comment: responsible Comment: for increased efficiency of digestion, prevention of neonate Comment: hypersensitivity to ingested proteins, and inhibition of Comment: gastric pathogens. Number of members: 56</p>
CAT	PDOC00093	Chloramphenicol acetyltransferase	<p>Chloramphenicol acetyltransferase (CAT) (EC 2.3.1.28) [1] catalyzes the acetyl-CoA dependent acetylation of chloramphenicol (Cm), an antibiotic which inhibits prokaryotic peptidyltransferase activity. Acetylation of Cm by CAT inactivates the antibiotic. A histidine residue, located in the C-terminal section of the enzyme, plays a central role in its catalytic mechanism. We</p>

868

			<p>derived a signature pattern from the region surrounding this active site residue.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern Q-[LIV]-H-H-[SA]-x(2)-D-G-[FY]-H [The second H is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE</p> <p>Note there is a second family of CAT [2], evolutionary unrelated to the main family described above. These CAT belong to the bacterial hexapeptide-repeat containing-transferases family (see <PDOC00094>).</p> <p>Last update November 1997 / Text revised.</p> <p>References [1] Shaw W V., Leslie A G W. Annu. Rev. Biophys. Chem. 20:363-386(1991)</p> <p>[2] Parent R., Roy P H. J Bacteriol. 174 2891-2897(1992).</p>
Cation_efflux		Cation efflux family	<p>Accession number. PF01545</p> <p>Definition: Cation efflux family</p> <p>Author: Bateman A</p> <p>Alignment method of seed Clustalw</p> <p>Source of seed members: Pfam-B_232 (release 4.0)</p> <p>Gathering cutoffs: -6 -6</p> <p>Trusted cutoffs: 6 90 6 90</p> <p>Noise cutoffs: -19 30 -19 30</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmlcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98361887</p> <p>Reference Title: Molecular characterization of a chromosomal determinant conferring resistance to zinc and cobalt ions in <i>Staphylococcus aureus</i>.</p> <p>Reference Author: Xiong A. Jayaswal RK,</p> <p>Reference Location: J Bacteriol 1998;180:4024-4029.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 96219090</p> <p>Reference Title: Cloning and sequence analysis of <i>czc</i> genes in <i>Alcaligenes</i> sp. strain CT14.</p> <p>Reference Author: Kunito T, Kusano T, Oyaizu H, Senoo K, Kanazawa S,</p> <p>Reference Author: Matsumoto S;</p> <p>Reference Location: Biosci Biotechnol Biochem 1996;60:699-704</p> <p>Database Reference: INTERPRO; IPR002524;</p> <p>Database reference: PFAM; PB038216;</p> <p>Comment: Members of this family are integral membrane proteins, that</p> <p>Comment: are found to increase tolerance to divalent metal ions such as cadmium, zinc, and cobalt These proteins are thought to</p> <p>Comment: be efflux pumps that remove these ions from cells</p> <p>Number of members: 59</p>
CBD_6		Cellulose binding domain	<p>Accession number: PF02018</p> <p>Definition: Cellulose binding domain</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Manual</p> <p>Source of seed members: Chris Ponting</p> <p>Gathering cutoffs: 19 0</p> <p>Trusted cutoffs: 19.10 19.10</p> <p>Noise cutoffs: 8.90 8.90</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmlcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97074498</p> <p>Reference Title: Structure of the N-terminal cellulose-binding domain of</p>

			<p>Reference Title: Cellulomonas fimi CenC determined by nuclear magnetic resonance spectroscopy.</p> <p>Reference Author: Johnson PE, Joshi MD, Tomme P, Kilburn DG, McIntosh LP;</p> <p>Reference Location: Biochemistry 1996;35:14381-14394.</p> <p>Database Reference: URL; http://www.ocms.ox.ac.uk/~ponting/methmb/example.html;</p> <p>Database Reference: SCOP; 1ulp; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: PDB; 1ulo ; 1; 149;</p> <p>Database Reference: PDB; 1ulp ; 1; 149;</p> <p>Database Reference: PDB; 1cx1 A; 2; 6;</p> <p>Database Reference: PDB; 1ulo ; 150; 152;</p> <p>Database Reference: PDB; 1ulp , 150, 152;</p> <p>Database Reference: PDB; 1cx1 A; 7; 151;</p> <p>Database reference: PFAMB; PB012497;</p> <p>Database reference: PFAMB; PB041237;</p> <p>Database reference: PFAMB; PB041605;</p> <p>Number of members: 76</p>
CBFD_NFYB_HMF	PDOC00578	CBF/NF-Y subunits signatures	<p>Diverse DNA binding proteins are known to bind the CCAAT box, a common cis-acting element found in the promoter and enhancer regions of a large number of genes in eukaryotes. Amongst these proteins is one known as the CCAAT-binding factor (CBF) or NF-Y [1] CBF is a heteromeric transcription factor that consists of two different components both needed for DNA-binding</p> <p>The HAP protein complex of yeast binds to the upstream activation site of cytochrome C iso-1 gene (CYC1) as well as other genes involved in mitochondrial electron transport and activates their expression. It also recognizes the sequence CCAAT and is structurally and evolutionary related to CBF.</p> <p>The first subunit of CBF, known as CBF-A or NF-YB in vertebrates, HAP3 in budding yeast and as php3 in fission yeast, is a protein of 116 to 210 amino-acid residues which contains a highly conserved central domain of about 90 residues. This domain seems to be involved in DNA-binding; we have developed a signature pattern from its central part.</p> <p>The second subunit of CBF, known as CBF-B or NF-YA in vertebrates. HAP2 in budding yeast and php2 in fission yeast, is a protein of 265 to 350 amino-acid residues which contains a highly conserved region of about 60 residues. This region, called the 'essential core' [2], seems to consist of two subdomains an N-terminal subunit-association domain and a C-terminal DNA recognition domain. We have developed a signature pattern from a section of the subunit-association domain.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-V-S-E-x-I-S-F-[LIVM]-T-[SG]-E-A-[SC]-[DE]-[KRQ]-C Sequences known to belong to this class detected by the pattern ALL CBF-A subunits. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern Y-V-N-A-K-Q-Y-x-R-I-L-K-R-R-x-A-R-A-K-L-E Sequences known to belong to this class detected by the pattern ALL CBF-B subunits. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1995 / Patterns and text revised.</p> <p>References [1] Li X.-Y., Mantovani R., Hooft van Huijsduijnen R., Andre I., Benoist C., Mathis D. Nucleic Acids Res. 20:1087-1091(1992).</p> <p>[2] Olesen J.T , Fikes J D., Guarente L. Mol. Cell. Biol. 11:611-619(1991).</p>

870

CbiX		CbiX	<p>Accession number: PF01903</p> <p>Definition: CbiX</p> <p>Author: Enright A, Ouzounis C, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Enright A</p> <p>Gathering cutoffs: -25 -25</p> <p>Trusted cutoffs: -23.10 -23.10</p> <p>Noise cutoffs: -35.10 -35.10</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98416126</p> <p>Reference Title: Cobalamin (vitamin B12) biosynthesis: identification and characterization of a <i>Bacillus megaterium</i> cobI operon.</p> <p>Reference Author: Raux E, Lanois A, Warren MJ, Rambach A, Thermes C,</p> <p>Reference Location: Biochem J 1998;335:159-166.</p> <p>Database Reference: INTERPRO: IPR002762,</p> <p>Database reference: PFAMB; PB040604;</p> <p>Database reference: PFAMB; PB040610;</p> <p>Database reference: PFAMB; PB041575;</p> <p>Comment: The function of CbiX is uncertain, however it is found</p> <p>Comment: in cobalamin biosynthesis operons and so may have a</p> <p>Comment: related function. Some CbiX proteins contain a striking</p> <p>Comment: histidine-rich region at their C-terminus, which suggests</p> <p>Comment: that it might be involved in metal chelation [1].</p> <p>Number of members: 6</p>
cellulase	PDOC00565	Glycosyl hydrolases family 5 signature	<p>The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1,2]. Fungi and bacteria produces a spectrum of cellulolytic enzymes (cellulases) and xylanases which, on the basis of sequence similarities, can be classified into families. One of these families is known as the cellulase family A [3] or as the glycosyl hydrolases family 5 [4,E1]. The enzymes which are currently known to belong to this family are listed below.</p> <ul style="list-style-type: none"> - Endoglucanases from various species and strains of <i>Bacillus</i> - <i>Butyrivibrio fibrisolvens</i> endoglucanases 1 (end1) and A (celA). - <i>Caldocellum saccharolyticum</i> bifunctional endoglucanase/exoglucanase (celB) <p>This protein consists of two domains, it is the C-terminal domain, which has endoglucanase activity, which belongs to this family</p> <ul style="list-style-type: none"> - <i>Clostridium acetobutylicum</i> endoglucanase (eglA). - <i>Clostridium cellulolyticum</i> endoglucanases A (celccA) and D (celccD). - <i>Clostridium cellulovorans</i> endoglucanase B (engB) and D (engD). - <i>Clostridium thermocellum</i> endoglucanases B (celB), C (celC), E (celE) G (celG) and H (celH). - <i>Erwinia chrysanthemi</i> endoglucanase Z (celZ). - <i>Fibrobacter succinogenes</i> endoglucanase 3 (cel-3) - <i>Pseudomonas fluorescens</i> endoglucanase C (celC) - <i>Pseudomonas solanacearum</i> endoglucanase (egl). - <i>Robillarda</i> strain Y-20 endoglucanase I - <i>Ruminococcus albus</i> endoglucanases I (EG-I), A (celA), and B (celB). - <i>Ruminococcus flavefaciens</i> cellodextrinase A (celA). - <i>Ruminococcus flavefaciens</i> endoglucanase E (celE) - <i>Streptomyces lividans</i> endoglucanase. - <i>Thermomonospora fusca</i> endoglucanase E-5 (celE) - <i>Trichoderma reesei</i> endoglucanase II (EGLII). - <i>Xanthomonas campestris</i> endoglucanase (engxcA). <p>As well as:</p> <ul style="list-style-type: none"> - Baker's yeast glucan 1,3-beta-glucosidase I/II (EC 3.2.1.58) (EXG1). - Baker's yeast glucan 1,3-beta-glucosidase 2 (EC 3.2.1.58) (EXG2) - Baker's yeast sporulation-specific glucan 1,3-beta-glucosidase (SPR1) - <i>Caldocellum saccharolyticum</i> beta-mannanase (EC 3.2.1.78) (manA). - Yeast hypothetical protein YBR056w - Yeast hypothetical protein YIR007w. <p>One of the conserved regions in these enzymes contains a conserved glutamic acid residue which is potentially involved [5] in the catalytic mechanism. We use this region as a signature pattern.</p>

			<p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIV]-[LIVMFYWGA](2)-[DNEQG]-[LIVMGST]-x-N-E-[PV]-[RHDNSTLIVFY] [E is a putative active site residue] Sequences known to belong to this class detected by the pattern ALL, except for Robillarda Y-20 endoglucanase I whose sequence is known to be incorrect and yeast YBR056w. Other sequence(s) detected in SWISS-PROT 22. Expert(s) to contact by email Henrissat B. bernie@afmb.cnrs-mrs.fr</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References [1] Beguín P. Annu. Rev. Microbiol. 44:219-248(1990).</p> <p>[2] Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J. Microbiol Rev 55:303-315(1991).</p> <p>[3] Henrissat B., Claeysens M., Tomme P., Lemesle L., Mornon J.-P. Gene 81:83-95(1989)</p> <p>[4] Henrissat B. Biochem. J. 280:309-316(1991).</p> <p>[5] Py B, Bortoli-German I., Harech J, Chippaux M, Barras F Protein Eng. 4:325-333(1991).</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?glycosid.txt</p>
CH	PDOC00019	Actinin-type actin-binding domain signatures	<p>Alpha-actinin is a F-actin cross-linking protein which is thought to anchor actin to a variety of intracellular structures [1] The actin-binding domain of alpha-actinin seems to reside in the first 250 residues of the protein. A similar actin-binding domain has been found in the N-terminal region of many different actin-binding proteins [2,3]:</p> <ul style="list-style-type: none"> - In the beta chain of spectrin (or fodrin). - In dystrophin, the protein defective in Duchenne muscular dystrophy (DMD) and which may play a role in anchoring the cytoskeleton to the plasma membrane. - In the slime mold gelation factor (or ABP-120). - In actin-binding protein ABP-280 (or filamin), a protein that link actin filaments to membrane glycoproteins. - In fimbrin (or plastin), an actin-bundling protein Fimbrin differs from the above proteins in that it contains two tandem copies of the actin-binding domain and that these copies are located in the C-terminal part of the protein. <p>We selected two conserved regions as signature patterns for this type of domain. The first of this region is located at the beginning of the domain, while the second one is located in the central section and has been shown to be essential for the binding of actin</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [EQ]-x(2)-[ATV]-[FY]-x(2)-W-x-N Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT 25.</p> <p>Consensus pattern [LIVM]-x-[SGN]-[LIVM]-[DAGHE]-[SAG]-x-[DNEAG]-[LIVM]-x-[DEAG]-x(4)-[LIVM]-x-[LM]-[SAG]-[LIVM]-[LIVMT]-W-x-[LIVM](2) Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE</p>

			<p>Last update November 1997 / Patterns and text revised.</p> <p>References</p> <p>[1] Schleicher M., Andre E., Harmann A., Noegel A.A. Dev. Genet. 9:521-530(1988).</p> <p>[2] Matsudaira P. Trends Biochem. Sci. 16 87-92(1991).</p> <p>[3] Dubreuil R.R. BioEssays 13:219-226(1991)</p>
chitinase_2	PDOC00839	Chitinases family 18 active site	<p>Chitinases (EC 3.2.1.14) [1] are enzymes that catalyze the hydrolysis of the beta-1,4-N-acetyl-D-glucosamine linkages in chitin polymers. From the view point of sequence similarity chitinases belong to either family 18 or 19 in the classification of glycosyl hydrolases [2,E1] Chitinases of family 18 (also known as classes III or V) groups a variety of proteins:</p> <p>a) Chitinases from:</p> <ul style="list-style-type: none"> - Prokaryotes such as Alteromonas, Bacillus, Serratia, Streptomyces, etc. - Plants such as Arabidopsis, cucumber, bean, tobacco, etc. - Fungi such as Aphanocladium, Rhizopus, Saccharomyces, etc. - Nematode (Brugia malayi). - Insects (Manduca sexta). - Baculoviruses (Autographa Californica Nuclear Polyhedrosis virus). <p>b) Other proteins:</p> <ul style="list-style-type: none"> - Hevamine, a rubber tree protein with chitinase and lysozyme activities. - Kluyveromyces lactis killer toxin alpha subunit, which acts as a chitinase. - Flavobacterium and Streptomyces endo-beta-N-acetylglucosaminidases (EC 3.2.1.96). - Mammalian di-N-acetylchitinase which is involved in the degradation of asparagine-linked glycoproteins. - Human cartilage glycoprotein Gp-39. - Jack bean concanavalin B (conB). a protein that has lost its catalytic activity. <p>Site directed mutagenesis experiments [3] and crystallographic data [4,5] have shown that a conserved glutamate is involved in the catalytic mechanism and probably acts as a proton donor. This glutamate is at the extremity of the best conserved region in these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVMFY]-[DN]-G-[LIVMF]-[DN]-[LIVMF]-[DN]-x-E [E is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for conB which has a Gln instead of the active site Glu.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Expert(s) to contact by email Neuhaus J.-M. jean-marc neuhaus@bota.unine.ch</p> <p>Henrissat B. bernie@afmb.cnrs-mrs.fr</p> <p>Last update November 1997 / Text revised.</p> <p>References</p> <p>[1] Flach J., Pilet P.-E., Jolles P. Experientia 48:701-716(1992).</p> <p>[2] Henrissat B. Biochem. J. 280:309-316(1991).</p> <p>[3]</p>

			<p>Watanabe T., Kohori K., Miyashita K., Fujii T., Sakai H., Uchida M., Tanaka H. J. Biol. Chem. 268:18567-18572(1993).</p> <p>[4] Perrakis A., Tews I., Dauter Z., Oppenheim A.B., Chet I., Wilson K.S., Vorgias C.E. Structure 2:1169-1180(1994).</p> <p>[5] van Scheltinga A.C.T., Kalk K.H., Beintema J.J., Dijkstra B.W. Structure 2:1181-1189(1994).</p> <p>[E1] http://www.expasy.ch/cgi-bi/lists?glycosid.txt</p>
Choline_kinase		Choline/ethanolamine kinase	<p>Accession number: PF01633 Definition: Choline/ethanolamine kinase Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1165 (release 4.1) Gathering cutoffs: 25 25 Trusted cutoffs: 242.90 242 90 Noise cutoffs: -85.90 -85 90 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98175949 Reference Title: Expression, purification, and characterization of choline kinase, product of the CKI gene from <i>Saccharomyces cerevisiae</i> Reference Author: Kim KH, Voelker DR, Flocco MT, Carman GM; Reference Location: J Biol Chem 1998;273:6844-6852. Database Reference: INTERPRO; IPR002573, Comment: Choline kinase catalyses the committed step in the synthesis of Comment: phosphatidylcholine by the CDP-choline pathway [1]. Number of members: 22</p>
Chorion		Chorion protein	<p>Accession number: PF01723 Definition: Chorion protein Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1914 (release 4.1) Gathering cutoffs: -46 -46 Trusted cutoffs: -43.70 -43 70 Noise cutoffs: -49.00 -49 00 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 95333194 Reference Title: Sequence analysis of a small early chorion gene subfamily Reference Title: interspersed within the late gene locus in <i>Bombyx mori</i>. Reference Author: Kravariti L, Lecanidou R, Rodakis GC; Reference Location: J Mol Evol 1995;41:24-33. Reference Number: [2] Reference Medline: 86313609 Reference Title: Evolution of the silk moth chorion gene superfamily: gene families CA and CB. Reference Author: Lecanidou R, Rodakis GC, Eickbush TH, Kafatos FC, Reference Location: Proc Natl Acad Sci U S A 1986;83 6514-6518. Database Reference: INTERPRO; IPR002635; Database reference: PFAM; PB009425; Comment: This family consists of the chorion superfamily proteins classes A, B, CA, CB and high-cysteine HCB from silk, gypsy and polyphemus moths. Comment: The chorion proteins make up the moths egg shell a complex Comment: extracellular structure [2]. Number of members: 35</p>
Chorismate_mut		Chorismate mutase	<p>Accession number: PF01817 Definition: Chorismate mutase Author: Bateman A</p>

874

			<p>Alignment method of seed: Manual Source of seed members: PSI-BLAST 1ecm Gathering cutoffs: 5 5 Trusted cutoffs: 5.10 5 10 Noise cutoffs: -19.90 -19.90 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 95062155 Reference Title: The crystal structure of allosteric chorismate mutase at 2.2-A resolution Reference Author: Xue Y, Lipscomb WN, Graf R, Schnappauf G, Braus G; Reference Location: Proc Natl Acad Sci U S A 1994;91:10814-10818. Reference Number: [2] Reference Medline: 98307941 Reference Title: Tyrosine and tryptophan act through the same binding site Reference Title: at the dimer interface of yeast chorismate mutase. Reference Author: Schnappauf G, Krappmann S, Braus GH; Reference Location: J Biol Chem 1998;273:17012-17017. Reference Number: [3] Reference Medline: 98165805 Reference Title: Chorismate mutase-prephenate dehydratase from Escherichia coli. Study of catalytic and regulatory domains using genetically engineered proteins. Reference Author: Zhang S, Pohnert G, Kongsaree P, Wilson DB, Clardy J, Reference Author: Ganem B; Reference Location: J Biol Chem 1998;273:6248-6253. Database Reference: SCOP: 1csm; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO, IPR002701; Database Reference: PDB, 1ecm B: 6; 89; Database Reference: PDB, 1ecm A: 5; 89; Database Reference: PDB, 1csm A: 133; 162; Database Reference: PDB, 3csm A: 133; 243; Database Reference: PDB, 3csm B: 133; 243; Database Reference: PDB, 4csm A: 133; 243; Database Reference: PDB, 4csm B: 133; 243; Database Reference: PDB, 5csm A: 133; 243; Database Reference: PDB, 2csm A: 133; 246; Comment: Chorismate mutase EC.5.4.99.5 catalyses the conversion of Comment: chorismate to prephenate in the pathway of tyrosine and Comment: phenylalanine biosynthesis. This enzyme is negatively Comment: regulated by tyrosine, tryptophan and phenylalanine [2,3] Number of members: 28</p>
CN_hydrolase	PDOC00712; PDOC00943	Nitrilases / cyanide hydratase signatures; Uncharacterized protein family UPF0012 signature	<p>Nitrilases (EC 3.5.5.1) are enzymes that convert nitriles into their corresponding acids and ammonia. They are widespread in microbes as well as in plants where they convert indole-3-acetonitrile to the hormone indole-3-acetic acid. A conserved cysteine has been shown [1,2] to be essential for enzyme activity; it seems to be involved in a nucleophilic attack on the nitrile carbon atom</p> <p>Cyanide hydratase (EC 4.2.1.66) converts HCN to formamide. In phytopathogenic fungi, it is used to avoid the toxic effect of cyanide released by wounded plants [3]. The sequence of cyanide hydrolase is evolutionary related to that of nitrilases.</p> <p>Yeast hypothetical proteins YIL164c and YIL165c also belong to this family.</p> <p>As signature patterns for these enzymes, we selected two conserved regions. The first is located in the N-terminal section while the second, which contains the active site cysteine, is located in the central section.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-x(2)-[LIVMFY](2)-x-[IF]-x-E-x(2)-[LIVM]-x-G-Y-P Sequences known to belong to this class detected by the pattern ALL.</p>

875

			<p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern G-[GAQ]-x(2)-C-[WA]-E-[NH]-x(2)-[PST]-[LIVMFYS]-x-[KR] [C is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1995 / Patterns and text revised.</p> <p>References [1] Kobayashi M., Izui H., Nagasawa T., Yamada H. Proc. Natl. Acad. Sci. U S A 90:247-251(1993)</p> <p>[2] Kobayashi M., Komeda H., Yanaka N., Nagasawa T., Yamada H J Biol. Chem 267:20746-20751(1992).</p> <p>[3] Wang P., Vanetten H D Biochem. Biophys. Res. Commun. 187:1048-1054(1992).</p> <p>The following uncharacterized proteins have been shown [1] to share regions of similarities:</p> <ul style="list-style-type: none"> - Yeast chromosome X hypothetical protein YJL126w. - Yeast chromosome XII hypothetical protein YLR351c. - Fission yeast hypothetical protein SpAC26A3.11 - Escherichia coli hypothetical protein ybeM. - Bacillus subtilis hypothetical protein yhcX. - Mycobacterium tuberculosis hypothetical protein MtCY20G9.06c. - Synechocystis strain PCC 6803 hypothetical protein sl0601. - A Pseudomonas fluorescens hypothetical protein in pqqF 5'region. - A Staphylococcus hypothetical protein in agr operon. <p>Except for yhcX which is larger, these are protein of about 30 to 35 Kd which contain, in their central section, a well conserved region centered on a cysteine residue.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [GTA]-x(2)-[IVT]-C-Y-D-[LIVM]-x-F-P-x(9)-G Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / First entry.</p> <p>References [1] Baroch A. Unpublished observations (1995).</p>
CorA		CorA-like Mg2+ transporter protein	<p>Accession number: PF01544</p> <p>Definition: CorA-like Mg2+ transporter protein</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_944 (release 4 0)</p> <p>Gathering cutoffs: -62 -62</p> <p>Trusted cutoffs: -5.90 -5.90</p> <p>Noise cutoffs: -86.20 -86 20</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98448512</p> <p>Reference Title: The CorA magnesium transporter gene family.</p> <p>Reference Author: Kehres DG, Lawyer CH, Maguire ME;</p> <p>Reference Location: Microb Comp Genomics 1998;3:151-169.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 99003207</p> <p>Reference Title: The CorA Mg2+ transport protein of Salmonella typhimurium.</p> <p>Reference Title: Mutagenesis of conserved residues in the third membrane</p> <p>Reference Title: domain identifies a Mg2+ pore.</p>

			<p>Reference Author. Smith RL, Szegedy MA, Kucharski LM, Walker C, Wiet RM, Reference Author. Redpath A, Kaczmarek MT, Maguire ME; Reference Location: J Biol Chem 1998;273:28663-28669. Database Reference INTERPRO; IPR002523; Database reference: PFAMB; PB041399; Comment: The CorA transport system is the primary Mg²⁺ influx system of Salmonella typhimurium and Escherichia coli. CorA is virtually ubiquitous in the Comment: Bacteria and Archaea There are also eukaryotic relatives of this protein Number of members: 25</p>
Cys_knot	PDOC00234	Glycoprotein hormones beta chain signatures	<p>Glycoprotein hormones [1,2] (or gonadotropins) are a family of proteins which include the mammalian hormones follitropin (FSH), lutropin (LSH), thyrotropin (TSH) and chorionic gonadotropin (CG), as well as at least two forms of fish gonadotropins. All these hormones consist of two glycosylated chains (alpha and beta). In mammalian gonadotropins, the alpha chain is identical in the four types of hormones but the beta chains, while homologous, are different.</p> <p>The beta chains are proteins of about 100 to 140 amino acid residues which contain twelve conserved cysteines all involved in disulfide bonds [3], as shown in the following schematic representation.</p> <p>xxx CxxxxxxxCxCxCxxxxxxxCxxxxxxxCxxxxxCxCxCxxxxxCxxxxxxxC xxx</p> <p>'C': conserved cysteine involved in a disulfide bond *': position of the patterns.</p> <p>We have developed two patterns for these hormones. The first one, located in the N-terminal section, is a region which has been said to be involved in the association between the two chains of the hormones. The second pattern consists of a cluster of five conserved cysteines in the C-terminal section.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-[STAGM]-G-[HFYL]-C-x-[ST] [The two C's are involved in disulfide bonds] Sequences known to belong to this class detected by the pattern ALL, except for rat beta-FSH which has Glu in position 2 of the pattern. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [PA]-V-A-x(2)-C-x-C-x(2)-C-x(4)-[STD]-[DEY]-C-x(6,8)-[PGSTAVM]-x(2)-C [The five C's are involved in disulfide bonds] Sequences known to belong to this class detected by the pattern ALL, except for 5 sequences. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Expert(s) to contact by email Laphorn A adrian@chem.gla.ac.uk</p> <p>Last update July 1998 / Patterns and text revised</p> <p>References [1] Pierce J G., Parsons T F. Annu. Rev. Biochem. 50 465-495(1981).</p> <p>[2] Stockell Hartree A , Renwick A.G.C. Biochem. J. 287:665-679(1992).</p>

			[3] Lapthorn A.J., Harris D C., Littlejohn A., Lustbader J W., Canfield R.E., Machin K J., Morgan F.J., Isaacs N.W Nature 369:455-461(1994).
cytochrome_b_C	PDOC00171	Cytochrome b/b6 signatures	<p>In the mitochondrion of eukaryotes and in aerobic prokaryotes, cytochrome b is a component of respiratory chain complex III (EC 1.10.2.2) - also known as the bc1 complex or ubiquinol-cytochrome c reductase. In plant chloroplasts and cyanobacteria, there is a analogous protein, cytochrome b6, a component of the plastoquinone-plastocyanin reductase (EC 1.10.99.1), also known as the b6f complex.</p> <p>Cytochrome b/b6 [1,2] is an integral membrane protein of approximately 400 amino acid residues that probably has 8 transmembrane segments. In plants and cyanobacteria, cytochrome b6 consists of two subunits encoded by the petB and petD genes. The sequence of petB is colinear with the N-terminal part of mitochondrial cytochrome b, while petD corresponds to the C-terminal part. Cytochrome b/b6 non-covalently binds two heme groups, known as b562 and b566.</p> <p>Four conserved histidine residues are postulated to be the ligands of the iron atoms of these two heme groups.</p> <p>Apart from regions around some of the histidine heme ligands, there are a few conserved regions in the sequence of b/b6. The best conserved of these regions includes an invariant P-E-W triplet which lies in the loop that separates the fifth and sixth transmembrane segments. It seems to be important for electron transfer at the ubiquinone redox site - called Qz or Qo (where o stands for outside) - located on the outer side of the membrane.</p> <p>A schematic representation of the structure of cytochrome b/b6 is shown below.</p> <pre> +---Fe-b562---+ +---Fe-b566--- + xxxxxxxxxxxxHxHxxxxxxxxxxxxHxHxxxxxxxxxxPEWxxxxxxxxxxxxxxxxxxxx <-----Cytochrome-b-----> <---Cytochrome-b6-petB-----><---Cytochrome-b6-petD-----> </pre> <p>We developed two signature patterns for cytochrome b/b6. The first includes the first conserved histidine of b/b6, which is a heme b562 ligand; the second includes the conserved PEW triplet.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [DENQ]-x(3)-G-[FYWMQ]-x-[LIVMF]-R-x(2)-H [H is a heme b562 ligand] Sequences known to belong to this class detected by the pattern ALL, except for 5 sequences Other sequence(s) detected in SWISS-PROT 15</p> <p>Consensus pattern P-[DE]-W-[FY]-[LFY](2) Sequences known to belong to this class detected by the pattern ALL, except for <i>Odocoileus hemionus</i> (mule deer) and <i>Paramecium tetraurelia</i> cytochrome b Other sequence(s) detected in SWISS-PROT 1.</p> <p>Last update November 1995 / Patterns and text revised.</p> <p>References [1] Howell N. J. Mol. Evol. 29:157-169(1989)</p> <p>[2] Esposti M.D., de Vries S., Crimi M., Ghelli A., Patarnello T., Meyer A. Biochim. Biophys. Acta 1143:243-271(1993).</p>
cytochrome_b_N	PDOC00171	Cytochrome b/b6 signatures	<p>In the mitochondrion of eukaryotes and in aerobic prokaryotes, cytochrome b is a component of respiratory chain complex III (EC 1.10.2.2) - also known as the bc1 complex or ubiquinol-cytochrome c reductase. In plant chloroplasts and cyanobacteria, there is a analogous protein, cytochrome b6, a component of the plastoquinone-plastocyanin reductase (EC 1.10.99.1), also known as the b6f complex.</p>

			<p>Cytochrome b/b6 [1,2] is an integral membrane protein of approximately 400 amino acid residues that probably has 8 transmembrane segments. In plants and cyanobacteria, cytochrome b6 consists of two subunits encoded by the petB and petD genes. The sequence of petB is colinear with the N-terminal part of mitochondrial cytochrome b, while petD corresponds to the C-terminal part. Cytochrome b/b6 non-covalently binds two heme groups, known as b562 and b566.</p> <p>Four conserved histidine residues are postulated to be the ligands of the iron atoms of these two heme groups.</p> <p>Apart from regions around some of the histidine heme ligands, there are a few conserved regions in the sequence of b/b6. The best conserved of these regions includes an invariant P-E-W triplet which lies in the loop that separates the fifth and sixth transmembrane segments. It seems to be important for electron transfer at the ubiquinone redox site - called Qz or Qo (where o stands for outside) - located on the outer side of the membrane.</p> <p>A schematic representation of the structure of cytochrome b/b6 is shown below.</p> <pre> +---Fe-b562---+ +---Fe-b566--- + xxxxxxxHxHxxxxxxxxHxHxxxxxxxxPEWxxxxxxxxxxxxxxx <-----Cytochrome-b-----> <---Cytochrome-b6-petB-----><---Cytochrome-b6-petD-----> </pre> <p>We developed two signature patterns for cytochrome b/b6. The first includes the first conserved histidine of b/b6, which is a heme b562 ligand; the second includes the conserved PEW triplet.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [DENQ]-x(3)-G-[FYWMQ]-x-[LIVMF]-R-x(2)-H [H is a heme b562 ligand] Sequences known to belong to this class detected by the pattern ALL, except for 5 sequences. Other sequence(s) detected in SWISS-PROT 15.</p> <p>Consensus pattern P-[DE]-W-[FY]-[LFY](2) Sequences known to belong to this class detected by the pattern ALL, except for <i>Odocoileus hemionus</i> (mule deer) and <i>Paramecium tetraurelia</i> cytochrome b. Other sequence(s) detected in SWISS-PROT 1.</p> <p>Last update November 1995 / Patterns and text revised.</p> <p>References [1] Howell N. J. Mol. Evol. 29:157-169(1989).</p> <p>[2] Esposti M.D., de Vries S., Crimi M., Ghelli A., Patarnello T., Meyer A. Biochim Biophys. Acta 1143 243-271(1993).</p>
cytochrome_c	PDOC00169	Cytochrome c family heme-binding site signature	<p>In proteins belonging to cytochrome c family [1], the heme group is covalently attached by thioether bonds to two conserved cysteine residues. The consensus sequence for this site is Cys-X-X-Cys-His and the histidine residue is one of the two axial ligands of the heme iron. This arrangement is shared by all proteins known to belong to cytochrome c family, which presently includes cytochromes c, c', c1 to c6, c550 to c556, cc3/Hmc, cytochrome f and reaction center cytochrome c.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-{CPWHF}-{CPWR}-C-H-{CFYW} Sequences known to belong to this class detected by the pattern ALL, except</p>

[illegible]

880

			<p>Consensus pattern [LIVM]-x-[AG]-[LIVMF](2)-N-x-T-x-D-S-F-x-D-x-[SG] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE</p> <p>Consensus pattern [GE]-[SA]-x-[LIVM](2)-D-[LIVM]-G-[GP]-x(2)-[STA]-x-P Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE. Last update November 1997 / Patterns and text revised. References [1] Slock J., Stahly D P., Han C.-Y., Six E.W , Crawford I.P. J. Bacteriol. 172:7211-7226(1990).</p> <p>[2] Volpes F., Dyer M., Scaife J G., Darby G., Stammers D.K., Delves C.J. Gene 112:213-218(1992).</p>
DHquinase_I	PDOC00788	Dehydroquinase class I active site	<p>3-dehydroquinase dehydratase (EC 4.2.1.10), or dehydroquinase, catalyzes the conversion of 3-dehydroquinase into 3-dehydroshikimate. It is the third step in the shikimate pathway for the biosynthesis of aromatic amino acids from chorismate. Two classes of dehydroquinases exist, known as types I and II. The best studied type I enzyme is from <i>Escherichia coli</i> (gene <i>aroD</i>) and related bacteria where it is a homodimeric protein of a chain of about 250 residues. In fungi, dehydroquinase is part of a multifunctional enzyme which catalyzes five consecutive steps in the shikimate pathway. In <i>aroD</i>, it has been shown [1] that a histidine is involved in the catalytic mechanism, we used the region around this residue as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern D-[LIVM]-[DE]-[LIVMN]-x(18.20)-[LIVM](2)-x-[SC]-[NHY]-H-[DN] [H is the active site residue] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE. Last update December 1999 / Pattern and text revised. References [1] Deka R.K., Kleanthous C., Coggins J.R. J. Biol. Chem. 267:22237-22242(1992).</p>
Diphthamide_syn		Putative diphthamide synthesis protein	<p>Accession number. PF01866 Definition: Putative diphthamide synthesis protein Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: 25 25 Trusted cutoffs: 44.90 44 90 Noise cutoffs: -174.70 -174.70 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 96183112 Reference Title: A cDNA from the ovarian cancer critical region of deletion on chromosome 17p13.3. Reference Author: Phillips NJ, Zeigler MR, Deaven LL; Reference Location: Cancer Lett 1996;102:85-90. Reference Number: [2] Reference Medline: 94010339 Reference Title: Diphthamide synthesis in <i>Saccharomyces cerevisiae</i>: structure of the DPH2 gene. Reference Author: Mattheakis LC, Sor F, Collier RJ; Reference Location: Gene 1993;132:149-154. Database Reference: INTERPRO; IPR002728; Comment: Swiss:Q16439 is a candidate tumour suppressor gene [1] DPH2 from Comment: yeast Swiss:P32461 [2], which confers resistance to diphtheria toxin Comment: has been found to be involved in diphthamide synthesis. Diphtheria</p>

881

			<p>Comment: toxin inhibits eukaryotic protein synthesis by ADP-ribosylating</p> <p>Comment: diphthamide, a posttranslationally modified histidine residue present</p> <p>Comment: in EF2. The exact function of the members of this family is</p> <p>Comment: unknown.</p> <p>Number of members: 12</p>
DLH		Dienelactone hydrolase family	<p>Accession number: PF01738</p> <p>Definition: Dienelactone hydrolase family</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_757 (release 4.2)</p> <p>Gathering cutoffs: 15 0</p> <p>Trusted cutoffs: 15 60 3.10</p> <p>Noise cutoffs: 14.40 14.40</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 90339491</p> <p>Reference Title: Refined structure of dienelactone hydrolase at 1.8 A.</p> <p>Reference Author: Pathak D, Ollis D;</p> <p>Reference Location: J Mol Biol 1990;214 497-525.</p> <p>Database Reference: SCOP; 1dn; fa, [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO, IPR002925:</p> <p>Database Reference: PDB; 1dn ; 16; 232;</p> <p>Database reference: PFAMB, PB004640:</p> <p>Database reference: PFAMB; PB041131:</p> <p>Database reference: PFAMB; PB041469:</p> <p>Number of members: 42</p>
DNA_mis_rep air	PDOC00057	DNA mismatch repair proteins mutL / hexB / PMS1 signature	<p>Mismatch repair contributes to the overall fidelity of DNA replication [1]. It involves the correction of mismatched base pairs that have been missed by the proofreading element of the DNA polymerase complex. The sequence of some proteins involved in mismatch repair in different organisms have been found to be evolutionary related. These proteins are:</p> <ul style="list-style-type: none"> - Escherichia coli and Salmonella typhimurium mutL protein [2]. MutL is required for dam-dependent methyl-directed DNA repair. - Streptococcus pneumoniae hexB protein [3]. The Hex system is nick directed. - Yeast proteins PMS1 and MLH1 [4]. - Human protein MLH1 [5] which is involved in a form of familial hereditary nonpolyposis colon cancer (HNPCC). <p>As a signature pattern for this class of mismatch repair proteins we selected a perfectly conserved heptapeptide which is located in the N-terminal section of these proteins</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-F-R-G-E-A-L</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE</p> <p>Last update</p> <p>November 1995 / Pattern and text revised</p> <p>References</p> <p>[1]</p> <p>Modrich P</p> <p>Annu. Rev. Biochem. 56:435-466(1987).</p> <p>[2]</p> <p>Mankovich J.A., McIntyre C.A., Walker G.C.</p> <p>J. Bacteriol. 171.5325-5331(1989).</p> <p>[3]</p> <p>Prudhomme M., Martin B., Mejean V., Claverys J.-P</p> <p>J. Bacteriol. 171.5332-5338(1989).</p> <p>[4]</p> <p>Prolla T.A., Christie D., Liskay R M.</p>

			Mol. Cell. Biol. 14:407-415(1994). [5] Bronner C.E., Baker S.M., Morrison P.T., Warren G., Smith L.G., Lescoe M.K., Kane M., Earibino C., Lipford J., Linblom A., Tannergard P., Bollag R.J., Godwin A.R., Ward D.C., Nordenskjold M., Fishel R., Kolodner R.D., Liskay R.M. Nature 368:258-261(1994).
DNA primase _S		DNA primase small subunit	Accession number: PF01896 Definition: DNA primase small subunit Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: 25 25 Trusted cutoffs: 198 40 198 40 Noise cutoffs: -120.80 -120.80 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 91219475 Reference Title: Mutations in conserved yeast DNA primase domains impair DNA Reference Title: replication in vivo. Reference Author: Francesconi S, Longhese MP, Piseri A, Santocanale C, Lucchini G, Plevani P: Reference Location: Proc Natl Acad Sci U S A 1991;88 3877-3881. Database Reference: INTERPRO; IPR002755, Comment: DNA primase synthesizes the RNA primers for the Okazaki Comment: fragments in lagging strand DNA synthesis. DNA primase Comment: is a heterodimer of large and small subunits. Number of members: 14
DnaB		DnaB-like helicase	Members of this family are comprise DNA replication enzymes which unwind the helix. Generally, such polypeptide are ATPases which move at the replication fork, disrupting hydrogen bonds. Such proteins are use for DNA replication in vivo and/or in vitro.
DnaJ C		DnaJ C terminal region	Accession number: PF01556 Definition: DnaJ C terminal region Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_342 (release 4.0) Gathering cutoffs: -24 -24 Trusted cutoffs: -22.60 -22 60 Noise cutoffs: -25.50 -25 50 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98308847 Reference Title: The J-domain family and the recruitment of chaperone power. Reference Author: Kelley WL; Reference Location: Trends Biochem Sci 1998;23:222-227. Database Reference: INTERPRO; IPR002939; Database reference: PFAMB; PB013976; Comment: This family consists of the C terminal region form the DnaJ Comment: protein. Although the function of this region is unknown, it Comment: is always found associated with DnaJ and DnaJ_CXXCXGXG. Comment: DnaJ is a chaperone associated with the Hsp70 heat-shock Comment: system involved in protein folding and renaturation after stress. Number of members: 116
DnaJ_CXXCXGXG	PDOC00553	dnaJ domains signatures and profile	The prokaryotic heat shock protein dnaJ interacts with the chaperone hsp70-like dnaK protein [1]. Structurally, the dnaJ protein consists of an N-terminal conserved domain (called 'J' domain) of about 70 amino acids, a glycine-rich region ('G' domain') of about 30 residues, a central domain containing four repeats of a CXXCXGXG motif ('CRR' domain) and a C-terminal region of 120 to 170 residues. Such a structure is shown in the following

			<p>schematic representation.</p> <pre> +-----+-----+-----+-----+ N-terminal Gly-R CXXCXGXG C-terminal +-----+-----+-----+-----+ </pre> <p>It has been shown [2] that the 'J' domain as well as the 'CRR' domain are also found in other prokaryotic and eukaryotic proteins which are listed below.</p> <p>a) Proteins containing both a 'J' and a 'CRR' domain:</p> <ul style="list-style-type: none"> - Yeast protein MAS5/YDJ1 which seems to be involved in mitochondrial protein import - Yeast protein MDJ1, involved in mitochondrial biogenesis and protein folding. - Yeast protein SCJ1, involved in protein sorting. - Yeast protein XDJ1. - Plants dnaJ homologs (from leek and cucumber). - Human HDJ2, a dnaJ homolog of unknown function. - Yeast hypothetical protein YNL077w. <p>b) Proteins containing a 'J' domain without a 'CRR' domain</p> <ul style="list-style-type: none"> - Rhizobium fredii nolC, a protein involved in cultivar-specific nodulation of soybean. - Escherichia coli cbpA [3], a protein that binds curved DNA. - Yeast protein SEC63/NPL1 important for protein assembly into the endoplasmic reticulum and the nucleus - Yeast protein SIS1, required for nuclear migration during mitosis. - Yeast protein CAJ1. - Yeast hypothetical protein YFR041c. - Yeast hypothetical protein YIR004w - Yeast hypothetical protein YJL162c. - Plasmodium falciparum ring-infected erythrocyte surface antigen (RESA) RESA, whose function is not known, is associated with the membrane skeleton of newly invaded erythrocytes. - Human HDJ1. - Human HSJ1, a neuronal protein. - Drosophila cysteine-string protein (csp). <p>We developed a signature pattern for the 'J' domain, based on conserved positions in the C-terminal half of this domain. We also developed a pattern for the 'CRR' domain, based on the first two copies of that motif. We also developed a profile for the 'J' domain.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [FY]-x(2)-[LIVMA]-x(3)-[FYWHNT]-[DENQSA]-x-L-x-[DN]-x(3)-[KR]-x(2)-[FYI] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 5.</p> <p>Consensus pattern C-[DEGSTHKR]-x-C-x-G-x-[GK]-[AGSDM]-x(2)-[GSNKR]-x(4,6)-C-x(2,3)-C-x-G-x-G Sequences known to belong to this class detected by the pattern ALL, except for yeast XDJ1. Other sequence(s) detected in SWISS-PROT 8.</p> <p>Sequences known to belong to this class detected by the profile ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so. Expert(s) to contact by email Kelley W. kelley@medecine.unige.ch</p> <p>Last update July 1998 / Patterns and text revised. References</p>
--	--	--	---

			<p>[1] Cyr D.M., Langer T., Douglas M.G. Trends Biochem. Sci. 19:176-181(1994).</p> <p>[2] Bork P., Sander C., Valencia A., Bukau B Trends Biochem. Sci. 17:129-129(1992).</p> <p>[3] Ueguchi C., Kaneda M., Yamada H., Mizuno T. Proc. Natl Acad. Sci. U.S.A. 91:1054-1058(1994).</p>
dNK		Deoxynucleoside kinase	<p>Accession number: PF01712 Definition: Deoxynucleoside kinase Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1744 (release 4.1) Gathering cutoffs: 25 25 Trusted cutoffs: 47 50 47.50 Noise cutoffs: -5.40 -5.40 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 97236800 Reference Title: Cloning of the cDNA and chromosome localization of the gene Reference Title: for human thymidine kinase 2 Reference Author: Johansson M. Karlsson A; Reference Location: J Biol Chem 1997;272:8454-8458. Reference Number: [2] Reference Medline: 96293511 Reference Title: Cloning and expression of human deoxyguanosine kinase cDNA. Reference Author: Johansson M, Karlsson A; Reference Location: Proc Natl Acad Sci U S A 1996;93:7258-7262. Database Reference: INTERPRO; IPR002624, Comment: This family consists of various deoxynucleoside kinases Comment: cytidine EC 2.7.1.74, guanosine EC:2.7.1.113, adenosine EC:2.7.1.76 Comment: and thymidine kinase EC:2.7.1.21 (which also phosphorylates deoxyuridine Comment: and deoxycytosine.) These enzymes catalyse the production of Comment: deoxynucleotide 5'-monophosphate from a deoxynucleoside Comment: Using ATP and yielding ADP in the process. Number of members: 20</p>
DUF125		Integral membrane protein DUF125	<p>Accession number: PF01988 Definition: Integral membrane protein DUF125 Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: -60 -60 Trusted cutoffs: -57.90 -57.90 Noise cutoffs: -64.60 -64 60 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 95028150 Reference Title: Sequence, mapping and disruption of CCC1, a gene that cross-complements the Ca(2+)-sensitive phenotype of csg1 mutants. Reference Title: Fu D, Beeler T, Dunn T; Reference Author: Fu D, Beeler T, Dunn T; Reference Location: Yeast 1994;10:515-521. Database Reference: INTERPRO; IPR002839; Comment: This family of predicted integral membrane proteins has no known Comment: function. However it does include Swiss:P47818 that may have a Comment: role in regulating calcium levels [1] Number of members: 7</p>

885

DUF25		Domain of unknown function DUF25	<p>Accession number: PF01641</p> <p>Definition: Domain of unknown function DUF25</p> <p>Author: Bateman A, Enwright A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1539 (release 4.1)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 151.80 151.80</p> <p>Noise cutoffs: 10.60 10.60</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 20076492</p> <p>Reference Title: Novel selenoproteins identified in silico and in vivo by using a conserved RNA structural motif.</p> <p>Reference Author: Lescure A, Gautheret D, Carbon P, Krol A;</p> <p>Reference Location: J Biol Chem 1999;274.38147-38154.</p> <p>Database Reference: INTERPRO; IPR002579;</p> <p>Comment: This domain has no known function. It is found associated</p> <p>Comment: with the peptide methionine sulfoxide reductase enzymatic domain PMSR The domain has two conserved cysteine and histidines that could suggest and zinc binding site.</p> <p>Comment: The final cysteine is found to be replaced by the rare amino</p> <p>Comment: acid selenocysteine in some members of the family [1].</p> <p>Number of members: 26</p>
DUF26		Domain of unknown function DUF26	<p>Accession number: PF01657</p> <p>Definition: Domain of unknown function DUF26</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_980 (release 4.1)</p> <p>Gathering cutoffs: -8 -8</p> <p>Trusted cutoffs: 6.50 1 40</p> <p>Noise cutoffs: -17.50 -17 50</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Database Reference: INTERPRO. IPR002902,</p> <p>Database reference: PFAM; PB005223;</p> <p>Comment: This domain has no known function It is found in serine/threonine</p> <p>Comment: kinases, associated with the Eukaryotic protein kinase domain</p> <p>Comment: pkinase. In the 33kDa secretory protein Swiss:082551</p> <p>Comment: this domain is duplicated. The domain contains four conserved</p> <p>Comment: cysteines</p> <p>Number of members: 25</p>
Dynein_light	PDOC00953	Dynein light chain type 1 signature	<p>Dynein is a multisubunit microtubule-dependent motor enzyme that acts as the force generating protein of eukaryotic cilia and flagella. The cytoplasmic isoform of dynein acts as a motor for the intracellular retrograde motility of vesicles and organelles along microtubules. Dynein is composed of a number of ATP-binding large subunits, intermediate size subunits and small subunits.</p> <p>Among the small subunits, there is a family [1,2] of highly conserved proteins which consist of:</p> <ul style="list-style-type: none"> - Chlamydomonas reinhardtii flagellar outer arm dynein 8 Kd and 11 Kd light chains - Higher eukaryotes cytoplasmic dynein light chain 1. - Yeast cytoplasmic dynein light chain 1 (gene DYN2 or SLC1). - Caenorhabditis elegans hypothetical dynein light chains M18.2 and T26A5.9. <p>These proteins are have from 89 to 120 amino acids. As a signature pattern, we selected a highly conserved region.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern H-x-l-x-G-[KR]-x-F-[GA]-S-x-V-[ST]-[HY]-E</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p>

886

			<p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / First entry</p> <p>References [1] King S M , Patel-King R.S. J. Biol. Chem 270 11445-11452(1995).</p> <p>[2] Dick T., Ray K., Salz H.K , Chia W. Mol. Cell. Biol. 16:1966-1977(1996).</p>
eIF5_eIF2B		Domain found in IF2B/IF5	<p>Accession number PF01873</p> <p>Definition: Domain found in IF2B/IF5</p> <p>Author: Enright A. Ouzounis C, Bateman A</p> <p>Alignment method of seed Clustalw</p> <p>Source of seed members: Enright A</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 233.00 233.00</p> <p>Noise cutoffs: -56.10 -56 10</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmlcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96060092</p> <p>Reference Title: Multidomain organization of eukaryotic guanine nucleotide</p> <p>Reference Title: exchange translation initiation factor eIF-2B subunits revealed by analysis of conserved sequence motifs.</p> <p>Reference Author: Koonin EV;</p> <p>Reference Location: Protein Sci 1995;4:1608-1617.</p> <p>Database Reference: INTERPRO. IPR002735,</p> <p>Comment: This family includes the N terminus of eIF-5</p> <p>Swiss.P55010, and</p> <p>Comment: the C terminus of eIF-2 beta Swiss:P20042 This region corresponds to the whole of the archaeobacterial eIF-2</p> <p>Comment: beta</p> <p>Comment: homolog. The region contains a putative zinc binding C4 finger.</p> <p>Number of members: 20</p>
eIF6		eIF-6 family	<p>This family comprises members exhibiting sequence identity to the eukaryotic translation initiation factor 6. Some members of this family are implicated in protein biosynthesis as a translation initiation factor by binding to the 60s ribosomal subunit and preventing its association with the 40s ribosomal subunit to form the 80s initiation complex. Such activity can play a role in maximal polysome formation and plays an important role in determining free 60s ribosomal subunit content. Polypeptides in this family can optimize amino acid and nitrogen content in a desired cell or organism. References describing eif6 family members and their biological activities include, for example, the following: Adams et al , Science 87:2185-2195(2000); Wood et al., J Biol. Chem 274:11653-11659(1999); and Si et al., Mol. Cell. Biol. 19:1416-1426(1999).</p>
ER	PDOC00992	Enhancer of rudimentary signature	<p>The Drosophila protein 'enhancer of rudimentary' (gene (e(r)) is a small protein of 104 residues whose function is not yet clear From an evolutionary point of view, it is highly conserved [1] and has been found to exist in probably all multicellular eukaryotic organisms. It has been proposed that this protein plays a role in the cell cycle</p> <p>As as signaure pattern, we selected a conserved region in the central part of the protein.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern Y-D-I-[SA]-x-L-[FY]-x-F-[IV]-D-x(3)-D-[LIV]-S</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / First entry.</p> <p>References [1] Gelsthorpe M., Pulumati M., McCallum C., Dang-Vu K., Tsubota S.I. Gene 186:189-195(1997).</p>

ER_lumen_re cept	PDOC00732	ER lumen protein retaining receptor signatures	<p>Proteins that reside in the lumen of the endoplasmic reticulum (ER) contain a C-terminal tetrapeptide (generally K-D-E-L or H-D-E-L) that serves as a signal for their retrieval (retrograde transport) from subsequent compartments of the secretory pathway. The signal is recognized by a receptor molecule that is believed to cycle between the cis side of the Golgi apparatus and the ER [1]. This protein is known as the ER lumen protein retaining receptor or also as the 'KDEL receptor'. It has been characterized in a variety of species, including fungi (gene ERD2), plants, Plasmodium, Drosophila and mammals. In mammals two highly related forms of the receptor are known.</p> <p>Structurally, the receptor is a protein of about 220 residues that seems to contain seven transmembrane regions [2]. The N-terminal part (3 residues) is oriented toward the lumen while the C-terminal tail (about 12 residues) is cytoplasmic. There are three luminal and three cytoplasmic loops.</p> <p>We developed two signature patterns for these receptors. The first pattern corresponds to the C-terminal half of the first cytoplasmic loop as well as most of the second transmembrane domain. The second pattern is a perfectly conserved decapeptide that corresponds to the central part of the fifth transmembrane domain.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-[LIV]-S-x-[KR]-x-[QH]-x-L-[FY]-x-[LIV](2)-[FYW]-x(2)-R- Y Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE</p> <p>Consensus pattern L-E-[SA]-V-A-I-[LM]-P-Q-[LI] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update December 1999 / Patterns and text revised.</p> <p>References [1] Pelham H.R.B. Curr. Opin. Cell Biol. 3:585-591(1991)</p> <p>[2] Townsend F M., Wilson D.W., Pelham H R.B EMBO J 12:2821-2829(1993).</p>
ETF_alpha	PDOC00583	Electron transfer flavoprotein alpha-subunit signature	<p>The electron transfer flavoprotein (ETF) [1,2] serves as a specific electron acceptor for various mitochondrial dehydrogenases. ETF transfers electrons to the main respiratory chain via ETF-ubiquinone oxidoreductase. ETF is an heterodimer that consist of an alpha and a beta subunit and which bind one molecule of FAD per dimer. A similar system also exists in some bacteria.</p> <p>The alpha subunit of ETF is a protein of about 32 Kd which is structurally related to the bacterial nitrogen fixation protein fixB which could play a role in a redox process and feed electrons to ferredoxin.</p> <p>Other related proteins are:</p> <ul style="list-style-type: none"> - Escherichia coli hypothetical protein ydiR. - Escherichia coli hypothetical protein ygcQ. <p>As a signature pattern for these proteins we selected a highly conserved region which is located in the C-terminal section.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LI]-Y-[LIVM]-[AT]-x-G-[IV]-[SD]-G-x-[IV]-Q-H-x(2)-G-x(6)-[IV]-x-A-[IV]-N Sequences known to belong to this class detected by the pattern ALL, except for ygcQ. Other sequence(s) detected in SWISS-PROT NONE</p> <p>Last update July 1998 / Text revised.</p>

888

			<p>References</p> <p>[1] Finocchiaro G , Ikeda Y., Ito M , Tanaka K. Prog Clin. Biol. Res. 321:637-652(1990).</p> <p>[2] Tsai M.H . Saier M.H. Jr. Res Microbiol. 146:397-404(1995).</p>
Euk_porin	PDOC00483	Eukaryotic mitochondrial porin signature	<p>The major protein of the outer mitochondrial membrane of eukaryotes is a porin that forms a voltage-dependent anion-selective channel (VDAC) that behaves as a general diffusion pore for small hydrophilic molecules [1 to 4]. The channel adopts an open conformation at low or zero membrane potential and a closed conformation at potentials above 30-40 mV</p> <p>This protein contains about 280 amino acids and its sequence is composed of between 12 to 16 beta-strands that span the mitochondrial outer membrane. Yeast contains two members of this family (genes POR1 and POR2), vertebrates have at least three members (genes VDAC1, VDAC2 and VDAC3) [5]</p> <p>As a signature pattern we selected a conserved region located at the C-terminal part of these proteins</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [YH]-x(2)-D-[SPCAD]-x-[STA]-x(3)-[TAG]-[KR]-[LIVMF]-[DNSTA]-[DNS]-x(4)-[GSTAN]-[LIVMA]-x-[LIVMY] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE Last update July 1999 / Pattern and text revised. References</p> <p>[1] Benz R Biochim. Biophys. Acta 1197 167-196(1994).</p> <p>[2] Manella C.A Trends Biochem. Sci. 17:315-320(1992).</p> <p>[3] Dihanich M. Experientia 46:146-153(1990)</p> <p>[4] Forte M., Guy H R , Mannella C.A. J. Bioenerg. Biomembr 19:341-350(1987).</p> <p>[5] Sampson M.J., Lovell R.S., Davison D.B., Craigen W J Genomics 36:192-196(1996).</p>
F_bP_aldolase	PDOC00523	Fructose-bisphosphate aldolase class-II signatures	<p>Fructose-bisphosphate aldolase (EC 4.1.2.13) [1,2] is a glycolytic enzyme that catalyzes the reversible aldol cleavage or condensation of fructose-1,6-bisphosphate into dihydroxyacetone-phosphate and glyceraldehyde 3-phosphate.</p> <p>There are two classes of fructose-bisphosphate aldolases with different catalytic mechanisms. Class-II aldolases [2], mainly found in prokaryotes and fungi, are homodimeric enzymes which require a divalent metal ion - generally zinc - for their activity.</p> <p>This family also includes the following proteins:</p> <ul style="list-style-type: none"> - Escherichia coli galactitol operon protein gatY which catalyzes the transformation of tagatose 1,6-bisphosphate into glyceraldehyde 3-phosphate and D-glyceraldehyde 3-phosphate. - Escherichia coli N-acetyl galactosamine operon protein agaY which catalyzes the same reaction as that of gatY. <p>As signature patterns for this class of enzyme, we selected two conserved</p>

			<p>regions The first pattern is located in the first half of the sequence and contains two histidine residues that have been shown [4] to be involved in binding a zinc ion. The second is located in the C-terminal section and contains clustered acidic residues and glycines.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [FYVMT]-x(1,3)-[LIVMH]-[APNT]-[LIVM]-x(1,2)-[LIVM]-H-x-D- H-[GACH] [The two H's are zinc ligands] Sequences known to belong to this class detected by the pattern ALL. except for <i>Mycoplasma pneumoniae</i> aldolase. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [LIVM]-E-x-E-[LIVM]-G-x(2)-[IGM]-[GSTA]-x-E Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update December 1999 / Pattern and text revised.</p> <p>References [1] Perham R N. Biochem. Soc. Trans 18:185-187(1990).</p> <p>[2] Marsh J.J , Lebherz H G Trends Biochem. Sci. 17:110-113(1992).</p> <p>[3] von der Osten C H Barbas C.F. III, Wong C.-H , Sinskey A J. Mol Microbiol. 3:1625-1637(1989)</p> <p>[4] Berry A., Marshall K E. FEBS Lett. 318 11-16(1993).</p>
FAA_hydrolase	Fumarylacetoacetate (FAA) hydrolase family	Accession number: PF01557 Definition: Fumarylacetoacetate (FAA) hydrolase family Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_641 (release 4 0) Gathering cutoffs: 25 25 Trusted cutoffs: 42.10 42.10 Noise cutoffs: -93 10 -93.10 HMM build command line hmmbuild -F HMM SEED HMM build command line hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97255958 Reference Title: Mutations in the fumarylacetoacetate hydrolase gene causing Reference Title: hereditary tyrosinemia type I: overview. Reference Author: St-Louis M, Tanguay RM; Reference Location: Hum Mutat 1997;9:291-299. Reference Number: [2] Reference Medline: 96125235 Reference Title: Molecular characterization of the 4-hydroxyphenylacetate catabolic pathway of <i>Escherichia coli</i> W: engineering a mobile aromatic degradative cluster. Reference Title: Reference Author: Prieto MA, Diaz E, Garcia JL, Reference Location: J Bacteriol 1996;178:111-120. Reference Number: [3] Reference Medline: 96016123 Reference Title: Fungal metabolic model for human type I hereditary tyrosinaemia Reference Title: Reference Author: Fernandez-Canon JM, Penalva MA; Reference Location: Proc Natl Acad Sci U S A 1995;92:9132-9136. Reference Number: [4] Reference Medline: 94039092 Reference Title. Purification, nucleotide sequence and some properties of a Reference Title: bifunctional isomerase/decarboxylase from the Reference Title: homoprotocatechuate degradative pathway of <i>Escherichia coli</i>	

890

			<p>Reference Title: C.</p> <p>Reference Author: Roper DJ, Cooper RA;</p> <p>Reference Location: Eur J Biochem 1993;217:575-580.</p> <p>Database reference: MIM; 276700;</p> <p>Database Reference: INTERPRO; IPR002529,</p> <p>Comment: This family consists of fumarylacetoacetate (FAA) hydrolase,</p> <p>Comment: or fumarylacetoacetate hydrolase (FAH) and it also includes</p> <p>Comment: HHDD isomerase/OPET decarboxylase from E. coli strain W.</p> <p>Comment: FAA is the last enzyme in the tyrosine catabolic pathway, it hydrolyses</p> <p>Comment: fumarylacetoacetate into fumarate and acetoacetate which then join the</p> <p>Comment: citric acid cycle [1]. Mutations in FAA cause type I tyrosinemia in humans</p> <p>Comment: this is an inherited disorder mainly affecting the liver leading to</p> <p>Comment: liver cirrhosis, hepatocellular carcinoma, renal tubular damages and</p> <p>Comment: neurologic crises amongst other symptoms [1] The enzymatic defect causes</p> <p>Comment: the toxic accumulation of phenylalanine/tyrosine catabolites [3].</p> <p>Comment: The E. coli W enzyme HHDD isomerase/OPET decarboxylase contains two</p> <p>Comment: copies of this domain and functions in fourth and fifth steps of the</p> <p>Comment: homoprotocatechuate pathway;</p> <p>Comment: here it decarboxylates OPET to HHDD and isomerizes this to OHED.</p> <p>Comment: The final products of this pathway are pyruvic acid and succinic</p> <p>Comment: semialdehyde.</p> <p>Number of members: 33</p>
FAD_binding		FAD binding domain	<p>Accession number: PF00667</p> <p>Definition: FAD binding domain</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_180 (release 2.1)</p> <p>Gathering cutoffs: 16.8 16.8</p> <p>Trusted cutoffs: 24.60 16.80</p> <p>Noise cutoffs: 13.50 15.90</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmbuild --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 95386502</p> <p>Reference Title: The flavin reductase activity of the flavoprotein component</p> <p>Reference Title: of sulfite reductase from Escherichia coli. A new model for</p> <p>Reference Title: the protein structure.</p> <p>Reference Author: Eschenbrenner M, Coves J, Fontecave M;</p> <p>Reference Location: J Biol Chem 1995;270:20550-20555.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 96049560</p> <p>Reference Title: NADPH-sulfite reductase flavoprotein from Escherichia coli:</p> <p>Reference Title: contribution to the flavin content and subunit interaction</p> <p>Reference Author: Eschenbrenner M, Coves J, Fontecave M;</p> <p>Reference Location: FEBS Lett 1995;374:82-84.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 94360001</p> <p>Reference Title: Dissection of NADPH-cytochrome P450 oxidoreductase into</p> <p>Reference Title: distinct functional domains.</p> <p>Reference Author: Smith GC, Tew DG, Wolf CR;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1994;91:8710-8714.</p> <p>Reference Number: [4]</p> <p>Reference Medline: 97385116</p> <p>Reference Title: Three-dimensional structure of NADPH-cytochrome P450</p> <p>Reference Title: reductase: prototype for FMN- and FAD-containing</p>

891

			<p>enzymes.</p> <p>Reference Author Wang M, Roberts DL, Paschke R, Shea TM, Masters BS, Kim JJ;</p> <p>Reference Location. Proc Natl Acad Sci U S A 1997;94:8411-8416.</p> <p>Database Reference: SCOP; 1amo; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference INTERPRO; IPR001709;</p> <p>Database Reference PDB; 1amo A; 274; 493;</p> <p>Database Reference PDB; 1amo B; 274; 493;</p> <p>Database Reference PDB, 1quf , 77; 120;</p> <p>Database reference: PFAMB; PB001390;</p> <p>Comment: This domain is found in sulfite reductase, NADPH cytochrome P450</p> <p>Comment: reductase and Nitric oxide synthase.</p> <p>Number of members: 87</p>
FAD_binding_3		FAD binding domain	<p>Accession number: PF01494</p> <p>Definition: FAD binding domain</p> <p>Author Bashton M. Bateman A</p> <p>Alignment method of seed. Clustalw</p> <p>Source of seed members: Pfam-B_549 (release 4.0)</p> <p>Gathering cutoffs: -7 -7</p> <p>Trusted cutoffs -6 20 -6 20</p> <p>Noise cutoffs: -7.90 -7.90</p> <p>HMM build command line hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 93028353</p> <p>Reference Title: Crystal structure of the reduced form of p-hydroxybenzoate</p> <p>Reference Title: hydroxylase refined at 2.3A resolution.</p> <p>Reference Author: Schreuder HA, van der Laan JM, Swarte MB, Kalk KH. Hol WG,</p> <p>Reference Author: Drenth J;</p> <p>Reference Location. Proteins 1992;14:178-190.</p> <p>Database Reference: SCOP; 2phh; fa, [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference INTERPRO, IPR002938;</p> <p>Database Reference PDB; 1pxa ; 5; 35;</p> <p>Database Reference PDB; 1bf3 ; 5; 139;</p> <p>Database Reference PDB, 1bgj , 5; 139;</p> <p>Database Reference PDB; 1bgn ; 5; 139;</p> <p>Database Reference PDB; 1bkx ; 5; 139;</p> <p>Database Reference PDB, 1cc4 A; 5; 139;</p> <p>Database Reference PDB, 1cc6 A; 5; 139;</p> <p>Database Reference PDB, 1cj2 A; 5; 139;</p> <p>Database Reference PDB; 1pbb ; 5; 139;</p> <p>Database Reference PDB; 1pbc ; 5; 139;</p> <p>Database Reference PDB, 1pbd , 5; 139;</p> <p>Database Reference PDB; 1pbe ; 5; 139;</p> <p>Database Reference PDB, 1pbf ; 5; 139;</p> <p>Database Reference PDB; 1pdh , 5; 139;</p> <p>Database Reference PDB; 2phh ; 5; 139;</p> <p>Database Reference PDB, 1cj3 A; 5; 139;</p> <p>Database Reference PDB; 1cj4 A; 5; 139;</p> <p>Database Reference PDB, 1phh , 5; 139;</p> <p>Database Reference PDB, 1d7l A; 5; 139;</p> <p>Database Reference PDB; 1dob ; 5; 139;</p> <p>Database Reference PDB; 1doc ; 5; 139;</p> <p>Database Reference PDB; 1dod ; 5; 139;</p> <p>Database Reference PDB, 1doe , 5; 139;</p> <p>Database Reference PDB; 1ius ; 5; 139;</p> <p>Database Reference PDB; 1iut ; 5; 139;</p> <p>Database Reference PDB; 1iuu ; 5; 139;</p> <p>Database Reference PDB; 1iuv ; 5; 139;</p> <p>Database Reference PDB, 1iuw ; 5; 139;</p> <p>Database Reference PDB; 1iux ; 5; 139;</p> <p>Database Reference PDB; 1foh A; 10; 151;</p> <p>Database Reference PDB; 1foh D; 10; 151;</p> <p>Database Reference PDB; 1foh B; 10; 151;</p> <p>Database Reference PDB; 1foh C; 10; 151;</p> <p>Database reference: PFAMB, PB040546;</p> <p>Comment: This domain is involved in FAD binding in a number of enzymes</p> <p>Number of members: 52</p>

892

FAD_binding _4	PDOC00674	Oxygen oxidoreductas es covalent FAD-binding site	<p>Some oxygen-dependent oxidoreductases are flavoproteins that contains a covalently bound FAD group which is attached to a histidine via an 8-alpha-(N3-histidyl)-riboflavin linkage. These proteins are:</p> <ul style="list-style-type: none"> - 6-hydroxy-D-nicotine oxidase (EC 1.5.3.6) (6-HDNO) [1], a bacterial enzyme that catalyzes the oxygen-dependent degradation of 6-hydroxynicotine into 6-hydroxypyrid-N-methylamine - Plant reticuline oxidase (EC 1.5.3.9) [2] (berberine-bridge-forming enzyme), an enzyme that catalyzes the oxidation of (S)-reticuline into (S)-scoulerine in the pathway leading to benzophenanthridine alkaloids. - L-gulonolactone oxidase (EC 1.1.3.8) (l-gulono-gamma-lactone oxidase) [3], a mammalian enzyme which catalyzes the oxidation of L-gulono-1,4-lactone to L-xylo-hexulonolactone which spontaneously isomerizes to L-ascorbate - D-arabinono-1,4-lactone oxidase (EC 1.1.3.24) (L-galactonolactone oxidase), a yeast enzyme involved in the biosynthesis of D-erythroascorbic acid [4] - Mitomycin radical oxidase [5], a bacterial protein involved in mitomycin resistance and that probably oxidizes the reduced form of mitomycins. - Rhodococcus fascians fasciation locus protein fas5. <p>The region around the histidine that binds the FAD group is conserved in these enzymes and can be used as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern P-x(10)-[DE]-[LIVM]-x(3)-[LIVM]-x(9)-[LIVM]-x(3)-[GSA]-[GST]-G-H [H is the FAD binding site] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update November 1997 / Text revised. EMBL/GenBank U40390 References [1] Brandsch R , Hinkkanen A.E., Mauch L., Nagursky H , Decker K. Eur. J Biochem. 167:315-320(1987). [2] Dittrich H., Kutchan T.M. Proc. Natl. Acad. Sci. U.S.A. 88:9969-9973(1991). [3] Koshizaka T., Nishikimi M., Ozawa T., Yagi K. J. Biol. Chem. 263:1619-1621(1988). [4] Huh W -K., Kim S.-T , Kim J.-Y., Hwang S.-W., Kang S.-O. [5] August P.R., Flickinger M C., Sherman D.H J. Bacteriol. 176:4448-4454(1994).</p>
fer2	PDOC00175; PDOC00642	2Fe-2S ferredoxins. iron-sulfur binding region signature; Adrenodoxin family, iron- sulfur binding region signature	<p>Ferredoxins [1] are a group of iron-sulfur proteins which mediate electron transfer in a wide variety of metabolic reactions. Ferredoxins can be divided into several subgroups depending upon the physiological nature of the iron sulfur cluster(s) and according to sequence similarities. One of these subgroups are the 2Fe-2S ferredoxins, which are proteins or domains of around one hundred amino acid residues that bind a single 2Fe-2S iron-sulfur cluster. The proteins that are known [2] to belong to this family are listed below.</p> <ul style="list-style-type: none"> - Ferredoxin from photosynthetic organisms; namely plants and algae where it is located in the chloroplast or cyanelle, and cyanobacteria. - Ferredoxin from archaebacteria of the Halobacterium genus. - Ferredoxin IV (gene pftA) and V (gene fdxD) from Rhodobacter capsulatus. - Ferredoxin in the toluene degradation operon (gene xylT) and naphthalene degradation operon (gene nahT) of Pseudomonas putida. - Hypothetical Escherichia coli protein yfaE. - The N-terminal domain of the bifunctional ferredoxin/ferredoxin reductase electron transfer component of the benzoate 1,2-dioxygenase complex (gene benC) from Acinetobacter calcoaceticus, the toluene 4-monooxygenase

			<p>complex (gene <i>tmoF</i>), the toluate 1,2-dioxygenase system (gene <i>xylZ</i>), and the xylene monooxygenase system (gene <i>xylA</i>) from <i>Pseudomonas</i>.</p> <ul style="list-style-type: none"> - The N-terminal domain of phenol hydroxylase protein <i>p5</i> (gene <i>dmpP</i>) from <i>Pseudomonas putida</i>. - The N-terminal domain of methane monooxygenase component C (gene <i>mmoC</i>) from <i>Methylococcus capsulatus</i>. - The C-terminal domain of the vanillate degradation pathway protein <i>vanB</i> in a <i>Pseudomonas</i> species - The N-terminal domain of bacterial fumarate reductase iron-sulfur protein (gene <i>frdB</i>). - The N-terminal domain of CDP-6-deoxy-3,4-glucose reductase (gene <i>ascD</i>) from <i>Yersinia pseudotuberculosis</i>. - The central domain of eukaryotic succinate dehydrogenase (ubiquinone) iron-sulfur protein. - The N-terminal domain of eukaryotic xanthine dehydrogenase. - The N-terminal domain of eukaryotic aldehyde oxidase. <p>In the 2Fe-2S ferredoxins, four cysteine residues bind the iron-sulfur cluster. Three of these cysteines are clustered together in the same region of the protein. Our signature pattern spans that iron-sulfur binding region.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-[C]-{C}-[GA]-{C}-C-[GAST]-{CPDEKRHFYW}-C [The three C's are 2Fe-2S ligands] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 15.</p> <p>Note in addition to the proteins listed above there are a number of other ferredoxin-like proteins that bind a 2Fe-2S cluster but which do not seem to be evolutionary related to this family. Among them are the ferredoxins from the adrenodoxin family (see <PDOC00642>) as well as the bacterial aromatic dioxygenase systems ferredoxin-like proteins such as <i>bnzC</i>, <i>ndoA</i>, and <i>todB</i>.</p> <p>Last update November 1997 / Text revised.</p> <p>References [1] Meyer J. <i>Trends Ecol. Evol.</i> 3:222-226(1988).</p> <p>[2] Harayama S., Polissi A., Rekik M. <i>FEBS Lett.</i> 285:85-88(1991)</p> <p>Ferredoxins [1] are a group of iron-sulfur proteins which mediate electron transfer in a wide variety of metabolic reactions. Ferredoxins can be divided into several subgroups depending upon the physiological nature of the iron sulfur cluster(s) and according to sequence similarities. One family of ferredoxins groups together the following proteins that all bind a single 2Fe-2S iron-sulfur cluster:</p> <ul style="list-style-type: none"> - Adrenodoxin (ADX) (adrenal ferredoxin), a vertebrate mitochondrial protein which transfers electrons from adrenodoxin reductase to cytochrome P450_{scc}, which is involved in cholesterol side chain cleavage. - Putidaredoxin (PTX), a <i>Pseudomonas putida</i> protein which transfers electrons from putidaredoxin reductase to cytochrome P450_{cam}, which is involved in the oxidation of camphor. - Terpredoxin [2], a <i>Pseudomonas</i> protein which transfers electrons from terpredoxin reductase to cytochrome P450_{terp}, which is involved in the oxidation of alpha-terpineol - Rhodocoxin [3], a <i>Rhodococcus</i> protein which transfers electrons from rhodocoxin reductase to cytochrome CYP116 (<i>thcB</i>), which is involved in the degradation of thiocarbamate herbicides. - <i>Escherichia coli</i> ferredoxin (gene <i>fdx</i>) [4] whose exact function is not yet known. - <i>Rhodobacter capsulatus</i> ferredoxin VI [5], which may transfer electrons to a yet uncharacterized oxygenase. - <i>Caulobacter crescentus</i> ferredoxin (gene <i>fdxB</i>) [6].
--	--	--	--

			<p>In these proteins, four cysteine residues bind the iron-sulfur cluster. Three of these cysteines are clustered together in the same region of the protein. Our signature pattern spans that iron-sulfur binding region.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-x(2)-[STAQ]-x-[STAMV]-C-[STA]-T-C-[HR] [The three C's are 2Fe-2S ligands] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT 1. Last update November 1995 / Pattern and text revised EMBL/Genbank: X51607. References [1] Meyer J Trends Ecol. Evol. 3:222-226(1988).</p> <p>[2] Peterson J A., Lu J.-Y., Geisselsoder J., Graham-Lorence S., Carmona C, Whitney F., Lorence M.C. J. Biol. Chem. 267 14193-14203(1992).</p> <p>[3] Nagy I., Schoofs G., Compennolle F., Proost P., Vanderleyden J., De Mot R. J. Bacteriol. 177 676-687(1995).</p> <p>[4] Ta D.T., Vickery L.E J. Biol. Chem. 267.11120-11125(1992).</p> <p>[5] Naud I., Vincon M., Garin J., Gaillard J., Forest E., Jouanneau Y. Eur. J. Biochem. 222 933-939(1994).</p> <p>[6] Amemiya K</p>
Ferric_reduct		Ferric reductase like transmembrane component	<p>Accession number. PF01794 Definition: Ferric reductase like transmembrane component Author: Bashton M, Bateman A Alignment method of seed T_Coffee Source of seed members. Pfam-B_728 (release 4.2) Gathering cutoffs: -122 -122 Trusted cutoffs: -34.80 -34.80 Noise cutoffs: -210.30 -210.30 HMM build command line. hmmbuild -F HMM SEED HMM build command line. hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 93309468 Reference Title: The fission yeast ferric reductase gene frp1+ is required for ferric iron uptake and encodes a protein that is homologous to the gp91-phox subunit of the human NADPH Reference Title: phagocyte oxidoreductase. Reference Author: Roman DG, Dancis A, Anderson GJ, Klausner RD, Reference Location: Mol Cell Biol 1993;13 4342-4350. Reference Number: [2] Reference Medline: 92294876 Reference Title: Cytochrome b558: the flavin-binding component of the phagocyte NADPH oxidase. Reference Title: phagocyte NADPH oxidase. Reference Author: Rotrosen D, Yeung CL, Leto TL, Malech HL, Kwong CH; Reference Location: Science 1992;256:1459-1462. Reference Number: [3] Reference Medline: 87258189 Reference Title: The glycoprotein encoded by the X-linked chronic granulomatous disease locus is a component of the neutrophil cytochrome b complex. Reference Title: neutrophil cytochrome b complex. Reference Author: Dinanuer MC, Orkin SH, Brown R, Jesaitis AJ, Parkos CA; Reference Location: Nature 1987;327:717-720.</p>

895

			<p>Reference Number: [4] Reference Medline: 87258190 Reference Title. The X-linked chronic granulomatous disease gene codes for Reference Title. the beta- chain of cytochrome b-245. Reference Author: Teahan C, Rowe P, Parker P, Totty N, Segal AW; Reference Location: Nature 1987;327:720-721 Database Reference INTERPRO; IPR002916; Comment: This family includes a common region in the transmembrane proteins Comment: mammalian cytochrome B-245 heavy chain (gp91-phox), ferric reductase Comment: transmembrane component in yeast and respiratory burst oxidase from Comment: mouse-ear cress. Comment: This may be a family of flavocytochromes capable of moving electrons Comment: across the plasma membrane [1]. Comment: The Frp1 protein Swiss:Q04800 from <i>S. pombe</i> is a ferric reductase Comment: component and is required for cell surface ferric reductase activity, Comment: mutants in <i>frp1</i> are deficient in ferric iron uptake [1]. Comment: Cytochrome B-245 heavy chain Swiss:P04839 is a FAD-dependent Comment: dehydrogenase it is also has electron transferase activity which reduces Comment: molecular oxygen to superoxide anion, a precursor in the production of Comment: microbicidal oxidants [2]. Comment: Mutations in the sequence of cytochrome B-245 heavy chain (gp91-phox) Comment: lead to the X-linked chronic granulomatous disease. The bacteriocidal Comment: ability of phagocytic cells is reduced and is characterized by the Comment: absence of a functional plasma membrane associated NADPH oxidase [3]. Comment: The chronic granulomatous disease gene codes for the beta chain of Comment: cytochrome B-245 and cytochrome B-245 is missing from patients with Comment: the disease [4]. Comment: The aligned region includes a potential FAD binding domain. Number of members: 34</p>
Flavi_NS5		Flavivirus RNA-directed RNA polymerase	<p>Accession number. PF00972 Definition: Flavivirus RNA-directed RNA polymerase Author. Finn RD, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_200 (release 3.0) Gathering cutoffs: 12 12 Trusted cutoffs. 16.00 16.00 Noise cutoffs: 8.50 8.50 HMM build command line hmmbuild -f HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 95159427 Reference Title: Phylogeny of TYU, SRE, and CFA virus: different evolutionary rates in the genus Flavivirus. Reference Author: Marin MS, Zanotto PM, Gritsun TS, Gould EA, Reference Location: Virology 1995;206:1133-1139. Reference Number: [2] Reference Medline: 96182933 Reference Title: Recombinant dengue type 1 virus NS5 protein expressed in Reference Title. <i>Escherichia coli</i> exhibits RNA-dependent RNA polymerase Reference Title: activity. Reference Author: Tan BH, Fu J, Sugrue RJ, Yap EH, Chan YC, Tan YH; Reference Location: Virology 1996;216:317-325 Reference Number: [3] Reference Medline: 93224895</p>

			<p>Reference Title: Computer-assisted identification of a putative methyltransferase domain in NS5 protein of flaviviruses and</p> <p>Reference Title: lambda 2 protein of reovirus</p> <p>Reference Author: Koonin EV;</p> <p>Reference Location: J Gen Virol 1993;74:733-740.</p> <p>Reference Number: [4]</p> <p>Reference Medline: 94094568</p> <p>Reference Title: Evolution and taxonomy of positive-strand RNA viruses. implications of comparative analysis of amino acid sequences.</p> <p>Reference Title: Koonin EV, Dolja VV;</p> <p>Reference Author: Crit Rev Biochem Mol Biol 1993;28:375-430.</p> <p>Reference Location: INTERPRO; IPR000208;</p> <p>Database Reference: Flaviviruses produce a polyprotein from the ssRNA genome.</p> <p>Comment: This protein is also known as NS5.</p> <p>Comment: This RNA-directed RNA polymerase possesses a number of short</p> <p>Comment: regions and motifs homologous to other RNA-directed RNA</p> <p>Comment: polymerases [2].</p> <p>Number of members: 159</p>
Fork_head	PDOC00564	Fork head domain signatures and profile	<p>It has been shown [1] that some eukaryotic transcription factors contain a conserved domain of about 100 amino-acid residues, called the fork head domain (but also known as a "winged helix"), which is involved in DNA-binding [2] Proteins known to contain this domain are listed below.</p> <ul style="list-style-type: none"> - Drosophila fork head protein (fkh). Fkh is probably a transcription factor that regulates the expression of genes involved in terminal development. - Drosophila protein crocodile (gene croc) [3], which is required for the establishment of head structures. - Drosophila proteins FD2, FD3, FD4, and FD5 - Drosophila proteins sloppy paired 1 and 2 (slp1 and slp2) involved in segmentation. - Bombyx mori silk gland factor-1 (SGF-1) which regulates transcription of the sericin-1 gene. - Mammalian transcriptional activators HNF-3-alpha, -beta, and -gamma. The HNF-3 proteins interact with the cis-acting regulatory regions of a number of liver genes. - Mammalian interleukin-enhancer binding factor (ILF). ILF binds to the purine-rich NFAT-like motifs in the HIV-1 LTR and the interleukin-2 promoter. ILF may be involved in both positive and negative regulation of important viral and cellular promoter elements. - Mammalian transcription factor BF-1 which plays an important role in the establishment of the regional subdivision of the developing brain and in the development of the telencephalon. - Human HTLF, a protein that binds to the purine-rich region in human T-cell leukemia virus long terminal repeat (HTLV-I LTR) - Mammalian transcription factors FREAC-1 (FKHL5, HFH-8), FREAC-2 (FKHL6), FREAC-3 (FKHL7, FKH-1), FREAC-4 (FKHL8), FREAC-5 (FKHL9, FKH-2, HFH-6), FREAC-6 (FKHL10 HFH-5), FREAC-7 (FKHL11), FREAC-8 (FKHL12, HFH-7), FKH-3, FKH-4, FKH-5, HFH-1 and HFH-4. - Human AFX1 which is involved in a chromosomal translocation that causes acute leukemia. - Human FKHR which is involved in a chromosomal translocation that causes rhabdomyosarcoma. - Xenopus XFKH1, a protein essential for normal axis formation. - Caenorhabditis elegans lin-31; involved in the regulation of vulval cell fates. - Yeast HCM1, a protein of unknown function. - Yeast FKH1. - Yeast FKH2 <p>The fork domain is highly conserved. We have developed two patterns for its detection. The first corresponds to the N-terminal section of the domain; the second is a heptapeptide located in the central section of the domain.</p>

			<p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [KR]-P-[PTQ]-[FYLVQH]-S-[FY]-x(2)-[LIVM]-x(3,4)-[AC]-[LIM]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for AFX1 and FKHR.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern W-[QKR]-[NS]-S-[LIV]-R-H</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Patterns and text revised.</p> <p>References</p> <p>[1] Weigel D., Jaeckle H. Cell 63:455-456(1990)</p> <p>[2] Clark K.L., Halay E.D., Lai E., Burley S K. Nature 364:412-420(1993)</p> <p>[3] Haecker U., Kaufmann E., Hartmann C , Juergens G , Knoechel W Jaeckle H EMBO J 14:5306-5317(1995).</p>
FtsJ		FtsJ cell division protein	<p>Accession number: PF01728</p> <p>Definition: FtsJ cell division protein</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1791 (release 4.1)</p> <p>Gathering cutoffs: -38 -38</p> <p>Trusted cutoffs: -20.90 -20 90</p> <p>Noise cutoffs: -56.70 -56.70</p> <p>HMM build command line. hmmbuild -F HMM SEED</p> <p>HMM build command line. hmmscalibrate --seed 0 HMM</p> <p>Reference Number. [1]</p> <p>Reference Medline: 93186701</p> <p>Reference Title: The Escherichia coli FtsH protein is a prokaryotic member of a protein family of putative ATPases involved in membrane functions, cell cycle control, and gene expression</p> <p>Reference Title: expression</p> <p>Reference Author Tomoyasu T, Yuki T, Morimura S, Mori H, Yamanaka K Niki H,</p> <p>Reference Author: Hiraga S, Ogura T,</p> <p>Reference Location. J Bacteriol 1993;175:1344-1351.</p> <p>Database Reference INTERPRO. IPR002877,</p> <p>Database reference: PFAMB; PB030182;</p> <p>Comment: This family consists of FtsJ from various bacterial and archaeal sources</p> <p>Comment: In E. coli FtsJ is not essential for growth but affects cell division [1].</p> <p>Number of members: 25</p>
FTSW_ROD A_SPOVE	PDOC00352	Cell cycle proteins ftsW / rodA / spoVE signature	<p>A number of prokaryotic proteins involved in cell cycle processes have been found [1,2] to be structurally related, these proteins are:</p> <ul style="list-style-type: none"> - Escherichia coli and related bacteria cell division protein ftsW. This protein plays a role in the stabilization of the ftsZ ring during cell division. - Escherichia coli and related bacteria rod shape-determining protein rodA (or mrdB). It is required for the expression of the enzymatic activity of PBP2, which is thought to participate in the synthesis of peptidoglycan during the initiation of cell elongation - Bacillus subtilis stage V sporulation protein E (spoVE). The exact function of spoVE in endospore formation is not known. - Bacillus subtilis hypothetical protein ylaO. - Bacillus subtilis hypothetical protein ywcF (ipa-42D). - Cyanophora paradoxa cyanelle ftsW homolog. This protein may be involved in the organelle division process. <p>All these proteins are hydrophobic integral membrane protein and contain about</p>

			<p>400 residues. We have selected the best conserved region, which is located in the C-terminal section, as a signature pattern for these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [NV]-x(5)-[GTR]-[LIVMA]-x-P-[PTLIVM]-x-G-[LIVM]-x(3)-[LIVMFVW](2)-S-[YSA]-G-G-[STN]-[SA]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References [1] Ikeda M., Sato T., Wachi M., Jung H.K., Ishino F., Kobayashi Y., Matsubashi M. J. Bacteriol. 171.6375-6378(1989).</p> <p>[2] Joris B., Dive G., Henriques A., Piggot P.J., Ghuyssen J.-M. Mol. Microbiol. 4:513-517(1990)</p>
Furin-like		Furin-like cysteine rich region	<p>Members of this family include receptors that mediate transmembrane signalling. These receptors can bind to a number of factors including amphiregulin, epidermal growth factor, gp30, heparin-binding egf, insulin, insulin-like growth factor I and II, neuregulins, transforming growth factor-alpha and, and vaccinia virus growth</p> <p>Signal transduction is mediated by catalytic activity of tyrosine kinase, such as ATP + A protein tyrosine = ADP + protein tyrosine phosphate. Typically, such signal transduction have been implicated in metabolic and developmental changes, including cell fate and differentiation. Examples include instruction of follicle cells to follow a dorsal pathway of development rather than the default ventral pathway. may also bind the spitz protein.</p> <p>References describing these family members and their biological activities:</p> <p>Abbot et al., J. Biol. Chem. 267:10759-10763(1992); Araki et al., J. Biol. Chem. 262:16186-16191(1987); Aroian et al., EMBO J. 13:360-366(1994); Aroian et al., Nature 348:693-699(1990); Barbetti et al., Diabetes 41:408-415(1992); Bargmann et al., Nature 319:226-230(1986); Cama et al., J. Biol. Chem. 268:8060-8069(1993); Cama et al., J. Clin. Endocrinol. Metab. 73:894-901(1991); Carrera et al., Hum. Mol. Genet. 2:1437-1441(1993); Clifford et al., Genetics 137 531-550(1994); Cocozza et al., Diabetes 41:521-526(1992); Cooke et al., Biochem. Biophys. Res. Commun. 177:1113-1120(1991); Coussens et al., Science 230:1132-1139(1985); Dickens et al., Biochem. Biophys. Res. Commun. 186:244-250(1992); Ebina et al., Cell 40:747-758(1985); Ebina et al., Proc. Natl. Acad. Sci. U.S.A. 84:704-708(1987); Ehsani et al., Genomics 15:426-429(1993); Elbein et al., Diabetes 42:429-434(1993); Elbein, Diabetes 38:737-743(1989); Fujita-Yamaguchi et al., Protein Seq. Data Anal. 1:3-6(1987); Gullick et al., EMBO J. 11:43-48(1992); Haruta et al., Diabetes 42:1837-1844(1993); Hubbard et al., EMBO J. 16:5572-5581(1997); Hubbard et al., Nature 372:746-754(1994); Iwanishi et al., Diabetologia 36:414-422(1993); Kadowaki et al., J. Clin. Invest. 86:254-264(1990); Kadowaki et al., Science 240:787-790(1988); Kim et al., Diabetologia 35:261-266(1992); Klinkhamer et al., EMBO J. 8:2503-2507(1989); Kusari et al., J. Biol. Chem. 266:5260-5267(1991); Lai et al., Neuron 6:691-704(1991); Lax et al., Mol. Cell Biol. 8:1970-1978(1988); Lebrun et al., J. Biol. Chem. 268:11272-11277(1993); Lee et al., Oncogene 8:3403-3410(1993); Lesokhin et al., Dev. Biol. 205:129-144(1999); Livneh et al., Cell 40:599-607(1985).</p> <p>Longo et al., Proc. Natl. Acad. Sci. U.S.A. 90:60-64(1993); McKeon et al., Mol. Endocrinol. 4:647-656(1990); Moller et al., J. Biol. Chem. 265:14979-14985(1990); Moller et al., Mol. Endocrinol. 4:1183-1191(1990); Odawara et al., Science 245:66-68(1989); Raz et al., Genetics 129:191-201(1991).</p> <p>Sakai et al., J. Mol. Biol. 256:548-555(1996); Schaeffer et al., Biochem. Biophys. Res. Commun. 189:650-653(1992); Schejter et al., Cell 46:1091-1101(1986); Seino et al., Biochem. Biophys. Res. Commun. 159:312-316(1989); Seino et al., Diabetes 39:123-128(1990); Semba et al., Proc. Natl. Acad. Sci. U.S.A. 82:6497-6501(1985); Shier et al., J. Biol. Chem. 264:14605-14608(1989); Taira et al., Science 245:63-66(1989); Tewari et al., J. Biol. Chem. 264:16238-16245(1989); Ullrich et al., Nature 313:756-761(1985).</p> <p>Ullrich et al., EMBO J. 5:2503-2512(1986); van der Vorm et al., Diabetologia 36:172-174(1993); van der Vorm et al., J. Biol. Chem. 267:66-71(1992); Wadsworth et al., Nature 314:178-180(1985); White et al., Cell 54:641-</p>

			649(1988); Xu et al., J. Biol. Chem. 265:18673-18681(1990); Yamamoto et al., Nature 319:230-234(1986); and Yoshimasa et al., Science 240:784-787(1988).
Galactosyl_T		Galactosyltransferase	<p>Accession number. PF01762</p> <p>Definition: Galactosyltransferase</p> <p>Author: Bashton M. Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_885 (release 4.2)</p> <p>Gathering cutoffs: -46 -46</p> <p>Trusted cutoffs: -43.90 -43.90</p> <p>Noise cutoffs: -49.80 -49.80</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98079080</p> <p>Reference Title: Cloning of a human</p> <p>Reference Title: UDP-galactose,2-acetamido-2-deoxy-D-glucose 3beta-galactosyltransferase catalyzing the formation of type 1 chains.</p> <p>Reference Author: Kolbinger F, Streiff MB, Katopodis AG;</p> <p>Reference Location: J Biol Chem 1998;273:433-440.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 98079027</p> <p>Reference Title: Genomic cloning and expression of three murine</p> <p>Reference Title: UDP-galactose beta-N- acetylglucosamine</p> <p>Reference Title: beta1,3-galactosyltransferase genes</p> <p>Reference Author: Hennet T, Dinter A, Kuhnert P, Mattu TS, Rudd PM, Berger</p> <p>Reference Author: EG,</p> <p>Reference Location: J Biol Chem 1998;273:58-65.</p> <p>Database Reference: INTERPRO; IPR002659.</p> <p>Database reference: PFAMB; PB005938;</p> <p>Database reference: PFAMB; PB012965;</p> <p>Comment: This family includes the galactosyltransferases</p> <p>Comment: UDP-galactose:2-acetamido-2-deoxy-D-glucose3beta-galactosyltransferase</p> <p>Comment: Swiss:O43825 [1] and UDP-Gal:beta-GlcNAc beta 1,3-galactosyltransferase</p> <p>Comment: Swiss:O54904 [2].</p> <p>Comment: Specific galactosyltransferases transfer galactose to GlcNAc terminal</p> <p>Comment: chains in the synthesis of the lacto-series oligosaccharides types 1</p> <p>Comment: and 2 [1].</p> <p>Number of members: 29</p>
G-alpha		G-protein alpha subunit	<p>Accession number PF00503</p> <p>Definition: G-protein alpha subunit</p> <p>Author: Finn RD</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_11 (release 1.0)</p> <p>Gathering cutoffs: 13.8 13.8</p> <p>Trusted cutoffs: 13.80 13.80</p> <p>Noise cutoffs: 9.70 12.70</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 94353239</p> <p>Reference Title: Structures of active conformations of Gi alpha 1 and the mechanism of GTP hydrolysis.</p> <p>Reference Title: Coleman DE, Berghuis AM, Lee E, Linder ME, Gilman AG,</p> <p>Reference Author: Sprang SR;</p> <p>Reference Location: Science 1994;265:1405-1412.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97004345</p> <p>Reference Title: How G proteins work: a continuing story.</p> <p>Reference Author: Coleman DE, Sprang SR;</p> <p>Reference Location: Trends Biochem Sci 1996;21:41-44.</p> <p>Database Reference: PRINTS; PR00318;</p> <p>Database Reference: SCOP; 1gia; fa, [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO, IPR001019;</p> <p>Database Reference: PDB; 1gia ; 34; 343;</p> <p>Database Reference: PDB; 1gil : 34; 343;</p>

900

			<p>Database Reference PDB; 1as0 ; 32: 344;</p> <p>Database Reference PDB; 1gfi ; 33: 345;</p> <p>Database Reference PDB; 1as2 ; 32: 346,</p> <p>Database Reference PDB; 1bh2 ; 32: 346,</p> <p>Database Reference PDB; 1cip A; 32: 347;</p> <p>Database Reference PDB; 1git ; 32: 348;</p> <p>Database Reference PDB; 1agr D; 11: 353;</p> <p>Database Reference PDB; 1gg2 A; 6: 348;</p> <p>Database Reference PDB; 1gp2 A, 6: 348;</p> <p>Database Reference PDB; 1bof ; 10: 353;</p> <p>Database Reference PDB; 1as3 ; 9: 353;</p> <p>Database Reference PDB; 1gdd ; 9: 353,</p> <p>Database Reference PDB; 1agr A; 6: 353;</p> <p>Database Reference PDB; 1tag ; 27: 340;</p> <p>Database Reference PDB; 1tad A; 27: 342,</p> <p>Database Reference PDB; 1tad B, 27: 342;</p> <p>Database Reference PDB; 1tnd B, 27: 342,</p> <p>Database Reference PDB; 1tnd C; 27: 342;</p> <p>Database Reference PDB; 1tad C; 27: 344;</p> <p>Database Reference PDB; 1tnd A, 27: 349,</p> <p>Database Reference PDB; 1cjk C; 39: 388;</p> <p>Database Reference PDB; 1cjt C; 39: 388;</p> <p>Database Reference PDB; 1cju C. 39: 388,</p> <p>Database Reference PDB; 1civ C; 39: 388,</p> <p>Database Reference PDB; 1azt A; 35: 391;</p> <p>Database Reference PDB; 1azt B; 35: 391;</p> <p>Database Reference PDB; 1azs C; 36: 393,</p> <p>Database reference: PFAMB; PB034080;</p> <p>Comment: G proteins couple receptors of extracellular signals to intracellular</p> <p>Comment: signaling pathways.</p> <p>Comment: The G protein alpha subunit binds guanyl nucleotide and is a weak</p> <p>Comment: GTPase.</p> <p>Number of members: 245</p>
GCV_H		Glycine cleavage H-protein	<p>Accession number. PF01597</p> <p>Definition: Glycine cleavage H-protein</p> <p>Author: Bateman A</p> <p>Alignment method of seed Clustalw</p> <p>Source of seed members Pfam-B_988 (release 4 1)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 27.90 27.90</p> <p>Noise cutoffs: -58.80 -58.80</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 94255425</p> <p>Reference Title: X-ray structure determination at 2.6-A resolution of a</p> <p>Reference Title: lipoate- containing protein: the H-protein of the glycine</p> <p>Reference Title: decarboxylase complex from pea leaves.</p> <p>Reference Author: Pares S Cohen-Addad C, Sieker L. Neuburger M. Douce R,</p> <p>Reference Location: Proc Natl Acad Sci U S A 1994;91:4850-4853.</p> <p>Database Reference: SCOP, 1hnp; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO, IPR002930;</p> <p>Database Reference: PDB; 1hpc A; 2: 127;</p> <p>Database Reference: PDB; 1hpc B; 2: 127;</p> <p>Database Reference: PDB; 1hnp ; 2: 127,</p> <p>Comment: This is a family of glycine cleavage H-proteins, part of the glycine</p> <p>Comment: cleavage multienzyme complex (GCV) found in bacteria</p> <p>Comment: and the mitochondria</p> <p>Comment: of eukaryotes. GCV catalyses the catabolism of glycine in eukaryotes.</p> <p>Comment: A lipoyl group is attached to a completely conserved lysine residue.</p> <p>Comment: The H protein shuttles the methylamine group of glycine from the</p> <p>Comment: P protein to the T protein.</p> <p>Number of members: 40</p>
GCV_T		Glycine cleavage T-	<p>Accession number: PF01571</p> <p>Definition: Glycine cleavage T-protein (aminomethyl transferase)</p>

901

		protein (aminomethyl transferase)	<p>Author: Bashton M. Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_933 (release 4 0)</p> <p>Gathering cutoffs: -146 -146</p> <p>Trusted cutoffs: -124.50 -124.50</p> <p>Noise cutoffs: -167.90 -167.90</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97199363</p> <p>Reference Title: Cloning, and molecular characterization of the GCV1 gene</p> <p>Reference Title: encoding the glycine cleavage T-protein from</p> <p>Saccharomyces</p> <p>Reference Title: cerevisiae.</p> <p>Reference Author: McNeil JB, Zhang F, Taylor BV, Sinclair DA, Pearlman RE,</p> <p>Reference Author: Bognar AL;</p> <p>Reference Location: Gene 1997;186:13-20.</p> <p>Database Reference: INTERPRO, IPR002536;</p> <p>Database reference: PFAMB: PB004229;</p> <p>Comment: This is a family of glycine cleavage T-proteins, part of the glycine</p> <p>Comment: cleavage multienzyme complex (GCV) found in bacteria</p> <p>and the mitochondria</p> <p>Comment: of eukaryotes. GCV catalyses the catabolism of glycine in eukaryotes</p> <p>Comment: The T-protein is an aminomethyl transferase</p> <p>Number of members: 27</p>
G-gamma	PDOC01002	G-protein gamma subunit profile	<p>Guanine nucleotide-binding proteins (G proteins) [1] act as intermediaries in the transduction of signals generated by transmembrane receptors. G proteins consist of three subunits (alpha, beta, and gamma). The alpha subunit binds to and hydrolyzes GTP; the functions of the beta and gamma subunits are less clear but they seem to be required for the replacement of GDP by GTP as well as for membrane anchoring and receptor recognition.</p> <p>The gamma subunits are small proteins (from 70 to 110 residues) that are bound to the membrane via a isoprenyl group (either a farnesyl or a geranyl-geranyl) covalently linked to their C-terminus. In mammals there are at least 12 different isoforms of gamma subunits.</p> <p>The <i>Caenorhabditis elegans</i> protein egl-10, which is a regulator of G-protein signalling, contains a G-protein gamma-like domain.</p> <p>We have developed a profile that spans the complete length of the gamma subunit.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Sequences known to belong to this class detected by the profile ALL, except for yeast and squid G-protein gamma.</p> <p>Other sequence(s) detected in SWISS-PROT NONE</p> <p>Expert(s) to contact by email</p> <p>Pennington S R. srpenn@liverpool.ac.uk</p> <p>Last update</p> <p>November 1997 / First entry.</p> <p>References</p> <p>[1]</p> <p>Pennington S.R.</p> <p>Protein Prof. 2:16-315(1995).</p>
glutaredoxin	PDOC00173	Glutaredoxin	<p>Glutaredoxin [1,2,3], also known as thioltransferase, is a small protein of approximately one hundred amino-acid residues. It functions as an electron carrier in the glutathione-dependent synthesis of deoxyribonucleotides by the enzyme ribonucleotide reductase. Like thioredoxin, which functions in a similar way, glutaredoxin possesses an active center disulfide bond. It exists in either a reduced or an oxidized form where the two cysteine residues are linked in an intramolecular disulfide bond.</p> <p>Glutaredoxin has been sequenced in a variety of species. On the basis of</p>

902

			<p>extensive sequence similarity, it has been proposed [4] that vaccinia protein O2L is most probably a glutaredoxin. Finally, it must be noted that phage T4 thioredoxin seems also to be evolutionary related.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVD]-[FYSA]-x(4)-C-[PV]-[FYWH]-C-x(2)-[TAV]-x(2,3)-[LIV] [The two C's form the redox-active bond] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note in position 5 of the pattern. all glutaredoxin sequences have Pro while T4 thioredoxin has Val. Last update December 1999 / Pattern and text revised. References [1] Gleason F.K., Holmgren A. FEMS Microbiol Rev 54:271-298(1988)</p> <p>[2] Holmgren A. Biochem. Soc. Trans 16:95-96(1988).</p> <p>[3] Holmgren A. J. Biol. Chem. 264:13963-13966(1989).</p> <p>[4] Johnson G P., Goebel S.J., Perkus M.E., Davis S.W., Winslow J.P., Paoletti E. Virology 181 378-381(1991).</p>
Glyco_hydro_1	PDOC00495	Glycosyl hydrolases family 1 signatures	<p>It has been shown [1 to 4] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family.</p> <ul style="list-style-type: none"> - Beta-glucosidases (EC 3.2.1.21) from various bacteria such as Agrobacterium strain ATCC 21400, Bacillus polymyxa, and Caldocellum saccharolyticum. - Two plants (clover) beta-glucosidases (EC 3.2.1.21). - Two different beta-galactosidases (EC 3.2.1.23) from the archaeobacteria Sulfolobus solfataricus (genes bgaS and lacS). - 6-phospho-beta-galactosidases (EC 3.2.1.85) from various bacteria such as Lactobacillus casei, Lactococcus lactis, and Staphylococcus aureus. - 6-phospho-beta-glucosidases (EC 3.2.1.86) from Escherichia coli (genes bgIB and ascB) and from Erwinia chrysanthemi (gene arbB) - Plants myrosinases (EC 3.2.3.1) (sinigrinase) (thioglucosidase). - Mammalian lactase-phlorizin hydrolase (LPH) (EC 3.2.1.108 / EC 3.2.1.62). <p>LPH, an integral membrane glycoprotein, is the enzyme that splits lactose in the small intestine. LPH is a large protein of about 1900 residues which contains four tandem repeats of a domain of about 450 residues which is evolutionary related to the above glycosyl hydrolases.</p> <p>One of the conserved regions in these enzymes is centered on a conserved glutamic acid residue which has been shown [5], in the beta-glucosidase from Agrobacterium, to be directly involved in glycosidic bond cleavage by acting as a nucleophile. We have used this region as a signature pattern. As a second signature pattern we selected a conserved region, found in the N-terminal extremity of these enzymes, this region also contains a glutamic acid residue</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVMFSTC]-[LIVFYS]-[LIV]-[LIVMST]-E-N-G-[LIVMFAR]-[CSAGN] [E is the active site residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 12.</p> <p>Note this pattern will pick up the last two domains of LPH; the first two domains, which are removed from the LPH precursor by proteolytic processing, have lost the active site glutamate and may therefore be inactive [4]</p> <p>Consensus pattern F-x-[FYWM]-[GSTA]-x-[GSTA]-x-[GSTA](2)-[FYNH]-[NQ]-x-</p>

			<p>E-x- [GSTA] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note this pattern will pick up the last three domains of LPH. Expert(s) to contact by email Henrissat B. bernie@afmb.cnrs-mrs.fr</p> <p>Last update November 1995 / Patterns and text revised. References [1] Henrissat B. Biochem. J. 280:309-316(1991).</p> <p>[2] Henrissat B. Protein Seq Data Anal 4:61-62(1991).</p> <p>[3] Gonzalez-Candelas L., Ramon D., Polaina J. Gene 95:31-38(1990).</p> <p>[4] El Hassouni M., Henrissat B., Chippaux M., Barras F. J. Bacteriol. 174:765-777(1992).</p> <p>[5] Withers S.G., Warren R.A.J., Street I.P., Rupitz K., Kempton J.B., Aebersold R. J. Am. Chem. Soc. 112:5887-5889(1990)</p>
Glyco_hydro_19	PDOC00620	Chitinases family 19 signatures	<p>Chitinases (EC 3.2.1.14) [1] are enzymes that catalyze the hydrolysis of the beta-1,4-N-acetyl-D-glucosamine linkages in chitin polymers. From the view point of sequence similarity chitinases belong to either family 18 or 19 in the classification of glycosyl hydrolases [2,E1]. Chitinases of family 19 (also known as classes IA or I and IB or II) are enzymes from plants that function in the defense against fungal and insect pathogens by destroying their chitin-containing cell wall. Class IA/I and IB/II enzymes differ in the presence (IA/I) or absence (IB/II) of a N-terminal chitin-binding domain (see the relevant entry <PDOC00025>). The catalytic domain of these enzymes consist of about 220 to 230 amino acid residues.</p> <p>As signature patterns we selected two highly conserved regions, the first one is located in the N-terminal section and contains one of the six cysteines which are conserved in most, if not all, of these chitinases and which is probably involved in a disulfide bond.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-x(4,5)-F-Y-[ST]-x(3)-[FY]-[LIVMF]-x-A-x(3)-[YF]-x(2)-F-[GSA] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE</p> <p>Consensus pattern [LIVM]-[GSA]-F-x-[STAG](2)-[LIVMFY]-W-[FY]-W-[LIVM] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Expert(s) to contact by email Neuhaus J.-M. jean-marc neuhaus@bota.unine.ch</p> <p>Henrissat B. bernie@afmb.cnrs-mrs.fr</p> <p>Last update November 1997 / Text revised. References [1] Flach J., Pilet P.-E., Jolles P. Experientia 48:701-716(1992).</p> <p>[2]</p>

			<p>Henrissat B. Biochem J. 280:309-316(1991)</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?glycosid.txt</p>
Glyco_hydro_3_C	PDOC00621	Glycosyl hydrolases family 3 active site	<p>It has been shown [1,2] that the following glycosyl hydrolases can be, on the basis of sequence similarities, classified into a single family:</p> <ul style="list-style-type: none"> - Beta glucosidases (EC 3.2.1.21) from the fungi <i>Aspergillus wentii</i> (A-3), <i>Hansenula anomala</i>, <i>Kluyveromyces fragilis</i>, <i>Saccharomycopsis fibuligera</i>, (BGL1 and BGL2), <i>Schizophyllum commune</i> and <i>Trichoderma reesei</i> (BGL1). - Beta glucosidases from the bacteria <i>Agrobacterium tumefaciens</i> (Cbg1), <i>Butyrivibrio fibrisolvens</i> (bglA), <i>Clostridium thermocellum</i> (bglB), <i>Escherichia coli</i> (bglX), <i>Erwinia chrysanthemi</i> (bgxA) and <i>Ruminococcus albus</i>. - <i>Alteromonas</i> strain O-7 beta-hexosaminidase A (EC 3.2.1.52). - <i>Bacillus subtilis</i> hypothetical protein yzbA. - <i>Escherichia coli</i> hypothetical protein ycfO and HI0959, the corresponding <i>Haemophilus influenzae</i> protein. <p>One of the conserved regions in these enzymes is centered on a conserved aspartic acid residue which has been shown [3], in <i>Aspergillus wentii</i> beta-glucosidase A3, to be implicated in the catalytic mechanism. We have used this region as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVM](2)-[KR]-x-[EQK]-x(4)-G-[LIVMFT]-[LIVT]-[LIVMF]-[ST]-D-x(2)-[SGADNI] [D is the active site residue] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE Expert(s) to contact by email Henrissat B. bernie@afmb.cnrs-mrs.fr</p> <p>Last update November 1997 / Pattern and text revised</p> <p>References [1] Henrissat B. Biochem. J. 280.309-316(1991).</p> <p>[2] Castle L.A., Smith K.D., Morris R.O. J. Bacteriol. 174:1478-1486(1992).</p> <p>[3] Bause E., Legler G. Biochim. Biophys. Acta 626:459-465(1980).</p>
Glyco_hydro_45	PDOC00877	Glycosyl hydrolases family 45 active site	<p>The microbial degradation of cellulose and xylans requires several types of enzymes such as endoglucanases (EC 3.2.1.4), cellobiohydrolases (EC 3.2.1.91) (exoglucanases), or xylanases (EC 3.2.1.8) [1,2]. Fungi and bacteria produces a spectrum of cellulolytic enzymes (cellulases) and xylanases which, on the basis of sequence similarities, can be classified into families. One of these families is known as the cellulase family K or as the glycosyl hydrolases family 45 [3,E1]. The enzymes which are currently known to belong to this family are listed below.</p> <ul style="list-style-type: none"> - Endoglucanase 5 from <i>Humicola insolens</i>. - Endoglucanase 5 from <i>Trichoderma reesei</i> (egl5) - Endoglucanase K from <i>Fusarium oxysporum</i>. - Endoglucanase B from <i>Pseudomonas fluorescens</i> (celB). - Endoglucanase 1 from <i>Ustilago maydis</i> (egl1). <p>The best conserved regions in these enzymes is located in the N-terminal section. It contains an aspartic acid residue which has been shown [4] to act as a nucleophile in the catalytic mechanism. We use this region as a signature pattern.</p>

			<p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [STA]-T-R-Y-[FYW]-D-x(5)-[CA] [The D is an active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE</p> <p>Expert(s) to contact by email</p> <p>Henrissat B. bernie@afmb.cnrs-mrs.fr</p> <p>Last update</p> <p>November 1997 / Pattern and text revised.</p> <p>References</p> <p>[1]</p> <p>Beguín P.</p> <p>Annu. Rev. Microbiol. 44:219-248(1990).</p> <p>[2]</p> <p>Gilkes N.R., Henrissat B., Kilburn D.G., Miller R.C. Jr., Warren R.A.J.</p> <p>Microbiol. Rev. 55:303-315(1991)</p> <p>[3]</p> <p>Henrissat B., Bairoch A.</p> <p>Biochem. J. 293:781-788(1993).</p> <p>[4]</p> <p>Davies G.J., Dodson G.G., Hubbard R.E., Tolley S.P., Dauter Z., Wilson K.S., Hjort C., Mikkelsen J.M., Rasmussen G., Schuelein M.</p> <p>Nature 365:362-364(1993)</p> <p>[E1]</p> <p>http://www.expasy.ch/cgi-bin/lists?glycosid.txt</p>
Glyco_hydro_47		Glycosyl hydrolase family 47	<p>Members of this family are alpha-mannosidases that catalyse the hydrolysis of the terminal 1,2-linked alpha-D-mannose residues in the oligo-mannose oligosaccharide Man(9)(GlcNAc)(2). These enzymes are capable of taking part in the glycosylation pathway and glycoprotein processing</p>
GTP_cyclohydrol	PDOC00672	GTP cyclohydrolase I signatures	<p>GTP cyclohydrolase I (EC 3.5.4.16) catalyzes the biosynthesis of formic acid and dihydroneopterin triphosphate from GTP. This reaction is the first step in the biosynthesis of tetrahydrofolate in prokaryotes, of tetrahydrobiopterin in vertebrates, and of pteridine-containing pigments in insects.</p> <p>GTP cyclohydrolase I is a protein of from 190 to 250 amino acid residues. The comparison of the sequence of the enzyme from bacterial and eukaryotic sources shows that the structure of this enzyme has been extremely well conserved throughout evolution [1].</p> <p>As signature patterns we selected two conserved regions. The first contains a perfectly conserved tetrapeptide which is part of the GTP-binding pocket [2]. the second region also contains conserved residues involved in GTP-binding</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [DEN]-[LIVM](2)-x(2)-[KRNQ]-[DEN]-[LIVM]-x(3)-[ST]-x-C-E-H-H</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [SA]-x-[RK]-x-Q-[LIVM]-Q-E-[RN]-[LI]-[TSN]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update</p> <p>July 1999 / Patterns and text revised</p> <p>References</p> <p>[1]</p> <p>Maier J., Witter K., Guetlich M., Ziegler I., Werner T., Ninnemann H.</p> <p>Biochem. Biophys. Res. Commun. 212:705-711(1995).</p> <p>[2]</p> <p>Nar H., Huber R., Meining W., Schmid C., Weinkauff S., Bacher A.</p>

906

			Structure 3:459-466(1995)
HCV_capsid		Hepatitis C virus capsid protein	<p>Family members include nucleocapsid proteins of the HCV. This virus family comprises a nucleocapsid covered by a lipoprotein envelope. The envelope consists of two proteins: protein M and glycoprotein E. The nucleocapsid is a complex of protein c and mRNA. Uses for these polypeptides include: immunological epitopes for vaccines; or as mRNA chaperone proteins to aid in processing or to prevent degradation</p> <p>References describing examples of these capsid polypeptides include: Chen et al, Virology 188:102-113(1992); and Okamoto et al., J. Gen. Virol. 72:2697-2704(1991)</p>
HD		HD domain	<p>Accession number: PF01966 Definition: HD domain Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: -1 -1 Trusted cutoffs: -0.50 -0.50 Noise cutoffs: -2.50 -2.50 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 99085258 Reference Title: The HD domain defines a new superfamily of metal-dependent phosphohydrolases. Reference Author: Aravind L, Koonin EV; Reference Location: Trends Biochem Sci 1998;23:469-472. Database Reference: INTERPRO, IPR002819; Database reference: PFAMB; PB005654; Database reference: PFAMB; PB006725; Database reference: PFAMB; PB009617; Database reference: PFAMB; PB012663; Database reference: PFAMB; PB035384; Database reference: PFAMB; PB040597; Comment: HD domains are metal dependent phosphohydrolases Number of members: 63</p>
HDV_ag		Hepatitis delta virus delta antigen	<p>Accession number: PF01517 Definition: Hepatitis delta virus delta antigen Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_808 (release 4.0) Gathering cutoffs: -8 -8 Trusted cutoffs: 23.30 23.30 Noise cutoffs: -40.50 -40.50 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 94065676 Reference Title: Characterization of RNA-binding domains of hepatitis delta antigen. Reference Author: Poisson F, Roingeard P, Baillou A, Dubois F, Bonelli F, Reference Author: Calogero RA, Goudeau A; Reference Location: J Gen Virol 1993;74:2473-2478 Reference Number: [2] Reference Medline: 98362586 Reference Title: Structural basis of the oligomerization of hepatitis delta antigen. Reference Author: Zuccola HJ, Rozzelle JE, Lemon SM, Erickson BW, Hogle JM; Reference Location: Structure 1998;6:821-830. Database Reference: SCOP; 1a92; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR002506; Database Reference: PDB; 1a92 A; 12; 23; Database Reference: PDB; 1a92 B; 12; 23; Database Reference: PDB; 1a92 C; 12; 23; Database Reference: PDB; 1a92 D; 12; 60; Database Reference: PDB; 1a92 A; 47; 60; Database Reference: PDB; 1a92 B; 47; 60; Database Reference: PDB; 1a92 C; 47; 60; Comment: The hepatitis delta virus (HDV) encodes a single protein,</p>

			<p>the</p> <p>Comment: hepatitis delta antigen (HDAg) The central region of this protein</p> <p>Comment: has been shown to bind RNA [1]. Several interactions are also</p> <p>Comment: mediated by a coiled-coil region at the N terminus of the protein [2].</p> <p>Number of members: 145</p>
hemolysinCa bind	PDOC00293	Hemolysin- type calcium- binding region signature	<p>Gram-negative bacteria produce a number of proteins which are secreted into the growth medium by a mechanism that does not require a cleaved N-terminal signal sequence. These proteins, while having different functions, seem [1] to share two properties they bind calcium and they contain a variable number of tandem repeats consisting of a nine amino acid motif rich in glycine, aspartic acid and asparagine. It has been shown [2] that such a domain is involved in the binding of calcium ions in a parallel beta roll structure. The proteins which are currently known to belong to this category are:</p> <ul style="list-style-type: none"> - Hemolysins from various species of bacteria. Bacterial hemolysins are exotoxins that attack blood cell membranes and cause cell rupture. The hemolysins which are known to contain such a domain are those from: <i>E. coli</i> (gene hlyA), <i>A. pleuropneumoniae</i> (gene appA), <i>A. actinomycetemcomitans</i> and <i>P. haemolytica</i> (leukotoxin) (gene lktA) - Cyclolysin from <i>Bordetella pertussis</i> (gene cyaA). A multifunctional protein which is both an adenylate cyclase and a hemolysin - Extracellular zinc proteases, serralyisin (EC 3.4.24.40) from <i>Serratia</i>, prtB and prtC from <i>Erwinia chrysanthemi</i> and aprA from <i>Pseudomonas aeruginosa</i>. - Nodulation protein nodO from <i>Rhizobium leguminosarum</i>. <p>We derived a signature pattern from conserved positions in the sequence of the calcium-binding domain</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern D-x-[L]-x(4)-G-x-D-x-[L]-x-G-G-x(3)-D Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note this pattern is found once in nodO and the extracellular proteases but up to 5 times in some hemolysin/cyclolysins.</p> <p>Last update October 1993 / Text revised.</p> <p>References [1] Economou A., Hamilton W D.O., Johnston A W.B , Downie J A. EMBO J. 9:349-354(1990).</p> <p>[2] Baumann U., Wu S ., Flaherty K M., McKay D B. EMBO J 12:3357-3364(1993).</p>
Herpes_alk_e xo		Herpesvirus alkaline exonuclease	<p>Accession number: PF01771</p> <p>Definition: Herpesvirus alkaline exonuclease</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_822 (release 4.2)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 318.00 318.00</p> <p>Noise cutoffs: -277.60 -277.60</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 85107093</p> <p>Reference Title: Studies on the herpes simplex virus alkaline nuclease:</p> <p>Reference Title: detection of type-common and type-specific epitopes on the</p> <p>Reference Title: enzyme.</p> <p>Reference Author: Banks LM, Halliburton IW, Purifoy DJ, Killington RA, Powell</p>

908

			<p>Reference Author. KL; Reference Location. J Gen Virol 1985;66:1-14. Database Reference INTERPRO; IPR001616; Comment: This family includes various alkaline exonucleases from members of the herpesviridae. Alkaline exonuclease Comment: appears to have an important role in the replication of herpes simplex virus [1]. Number of members: 23</p>
Herpes_gI		Alphaherpesvirus glycoprotein I	<p>Accession number: PF01688 Definition: Alphaherpesvirus glycoprotein I Author Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1222 (release 4.1) Gathering cutoffs: 25 25 Trusted cutoffs: 157.20 157.20 Noise cutoffs: -126.70 -126.70 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number. [1] Reference Medline: 96357074 Reference Title. Biosynthesis of glycoproteins E and I of feline herpesvirus: gE-gI interaction is required for intracellular transport. Reference Author: Mijnes JD, van der Horst LM, van Anken E, Horzinek MC, Reference Author. Rottier PJ, de Groot RJ; Reference Location: J Virol 1996;70:5466-5475 Reference Number. [2] Reference Medline: 94267406 Reference Title. Identification of the feline herpesvirus type 1 (FHV-1) genes encoding glycoproteins G, D, I and E: expression of Reference Title. FHV-1 glycoprotein D in vaccinia and raccoon poxviruses. Reference Author. Spatz SJ, Rota PA, Maes RK; Reference Location. J Gen Virol 1994;75:1235-1244 Reference Number. [3] Reference Medline: 94267879 Reference Title. Unusual phosphorylation sequence in the gpIV (gI) component Reference Title. of the varicella-zoster virus gpI-gpIV glycoprotein complex (VZV gE-gI complex) Reference Author: Yao Z, Grose C; Reference Location: J Virol 1994;68:4204-4211. Database Reference INTERPRO; IPR002874; Comment: This family consists of glycoprotein I from various members of the Comment: alphaherpesvirinae these include herpesvirus, varicella- zoster virus Comment: and pseudorabies virus. Glycoprotein I (gI) is important during natural Comment: infection. mutants lacking gI produce smaller lesions at the site of Comment: infection and show reduced neuronal spread [1] gI forms a heterodimeric Comment: complex with gE; this complex displays Fc receptor activity (binds to Comment: the Fc region of immunoglobulin) [1] Glycoproteins are also important Comment: in the production of virus-neutralizing antibodies and cell mediated Comment: immunity [2]. The alphaherpesvirinae have a dsDNA genome and have no Comment: RNA stage during viral replication. Number of members: 22</p>
Herpes_glycop_D		Herpesvirus glycoprotein M	<p>Accession number: PF01528 Definition: Herpesvirus glycoprotein M Author Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_929 (release 4.0) Gathering cutoffs: 25 25 Trusted cutoffs: 197.30 197.30 Noise cutoffs: -229.70 -229.70</p>

			<p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96357105</p> <p>Reference Title: Identification and characterization of pseudorabies virus glycoprotein gM as a nonessential virion component.</p> <p>Reference Author: Dijkstra JM, Visser N, Mettenleiter TC, Klupp BG;</p> <p>Reference Location: J Virol 1996;70:5684-5688.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 95381611</p> <p>Reference Title: Identification and molecular characterization of the murine cytomegalovirus homolog of the human cytomegalovirus UL100</p> <p>Reference Title: gene</p> <p>Reference Author: Li W, Eidman K, Gehrz RC, Kari B;</p> <p>Reference Location: Virus Res 1995;36:163-175.</p> <p>Database Reference: INTERPRO: IPR000785;</p> <p>Comment: The herpesvirus glycoprotein M (gM) is an integral membrane protein</p> <p>Comment: predicted to contain 8 transmembrane segments [2].</p> <p>Glycoprotein M is</p> <p>Comment: not essential for viral replication [1].</p> <p>Number of members: 24</p>
HesB-like	PDOC00887	Hypothetical hesB/yadR/yf hF family signature	<p>The following uncharacterized proteins have been shown [1] to share regions of similarities:</p> <ul style="list-style-type: none"> - Anabaena and related cyanobacteria protein hesB which may be required for nitrogen fixation. - Escherichia coli hypothetical protein yadR and HI1723, the corresponding Haemophilus influenzae protein. - Escherichia coli hypothetical protein ydiC. - Escherichia coli hypothetical protein yfhF and HI0376, the corresponding Haemophilus influenzae protein. - Mycobacterium tuberculosis hypothetical protein Rv2204c. - Synechocystis strain PCC 6803 hypothetical protein slr1417 - Synechocystis strain PCC 6803 hypothetical protein slr1565 - A hypothetical protein in the nifU 5' region of many nitrogen fixing bacteria. - Porphyra purpurea chloroplast hypothetical protein in apcF-rps4 intergenic region. - Yeast hypothetical protein YLL027W. - Yeast hypothetical protein YPR067W. <p>These are small proteins (106 to 135 amino-acid residues in bacteria, about 200 residues in fungi) that contain a number of conserved regions. The most noteworthy of these regions is located in the C-terminal extremity, it contains two conserved cysteines. We have used this region as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern F-x-[LIVMFY]-x-N-[PG]-[NSKQ]-x(4)-C-x-C-[GS]-x-S-F</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update</p> <p>December 1999 / Pattern and text revised.</p> <p>References</p> <p>[1]</p> <p>Barroch A., Rudd K.E</p> <p>Unpublished observations (1995).</p>
HisG	PDOC01020	ATP phosphoribosyltransferase signature	<p>ATP phosphoribosyltransferase (EC 2.4.2.17) is the enzyme that catalyzes the first step in the biosynthesis of histidine in bacteria, fungi and plants. It is a protein of about 23 to 32 Kd. As a signature pattern we selected a region located in the C-terminal part of this enzyme.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern E-x(5)-G-x-[SAG]-x(2)-[IV]-x-D-[LIV]-x(2)-[ST]-G-x-T-[LM]</p>

910

			<p>Sequences known to belong to this class detected by the pattern ALL</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update July 1998 / First entry</p>
histone	PDOC00045 PDOC00046 PDOC00287 PDOC00308	Histone H2A signature, Histone H4 signature, Histone H3 signature; Histone H2B signature	<p>Histone H2A is one of the four histones, along with H2B, H3 and H4, which forms the eukaryotic nucleosome core. Using alignments of histone H2A sequences [1,2,E1] we selected, as a signature pattern, a conserved region in the N-terminal part of H2A. This region is conserved both in classical S-phase regulated H2A's and in variant histone H2A's which are synthesized throughout the cell cycle.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [AC]-G-L-x-F-P-V</p> <p>Sequences known to belong to this class detected by the pattern ALL</p> <p>Other sequence(s) detected in SWISS-PROT 2.</p> <p>Last update November 1995 / Pattern and text revised.</p> <p>References</p> <p>[1] Wells D.E., Brown D Nucleic Acids Res. 19:2173-2188(1991).</p> <p>[2] Thatcher T H., Gorovsky M A Nucleic Acids Res. 22:174-179(1994).</p> <p>[E1] http://www.ncbi.nlm.nih.gov/Baxevani/HISTONES/index.html</p> <p>Histone H4 is one of the four histones, along with H2A, H2B and H3, which forms the eukaryotic nucleosome core. Along with H3, it plays a central role in nucleosome formation. The sequence of histone H4 has remained almost invariant in more than 2 billion years of evolution [1,E1]. The region we use as a signature pattern is a pentapeptide found in positions 14 to 18 of all H4 sequences. It contains a lysine residue which is often acetylated [2] and a histidine residue which is implicated in DNA-binding [3].</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-A-K-R-H</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT 1.</p> <p>Last update November 1995 / Text revised.</p> <p>References</p> <p>[1] Thatcher T.H., Gorovsky M.A. Nucleic Acids Res. 22 174-179(1994).</p> <p>[2] Doenecke D., Gallwitz D. Mol. Cell. Biochem. 44:113-128(1982).</p> <p>[3] Ebraldise K.K., Grachev S.A., Mirzabekov A.D. Nature 331:365-367(1988).</p> <p>[E1] http://www.ncbi.nlm.nih.gov/Baxevani/HISTONES/index.html</p> <p>Histone H3 is one of the four histones, along with H2A, H2B and H4, which forms the eukaryotic nucleosome core. It is a highly conserved protein of 135 amino acid residues [1,2,E1].</p> <p>The following proteins have been found to contain a C-terminal H3-like domain:</p>

911

		<p>- Mammalian centromeric protein CENP-A [3]. Could act as a core histone necessary for the assembly of centromeres.</p> <p>- Yeast chromatin-associated protein CSE4 [4].</p> <p>- Caenorhabditis elegans chromosome III encodes two highly related proteins (F54C8.2 and F58A4.3) whose C-terminal section is evolutionary related to the last 100 residues of H3. The function of these proteins is not yet known.</p> <p>We developed two signature patterns. The first one corresponds to a perfectly conserved heptapeptide in the N-terminal part of H3. The second one is derived from a conserved region in the central section of H3.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern K-A-P-R-K-Q-L Sequences known to belong to this class detected by the pattern ALL, except for the H3-like proteins and some protozoan H3. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern P-F-x-[RA]-L-[VA]-[KRQ]-[DEG]-[IV] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE Last update November 1997 / Patterns and text revised.</p> <p>References [1] Wells D.E., Brown D. Nucleic Acids Res. 19:2173-2188(1991).</p> <p>[2] Thatcher T H., Gorovsky M.A. Nucleic Acids Res 22:174-179(1994).</p> <p>[3] Sullivan K.F., Hechenberger M., Masri K. J. Cell Biol. 127:581-592(1994)</p> <p>[4] Stoler S , Keith K.C., Curnick K.E., Fitzgerald-Hayes M. Genes Dev. 9:573-586(1995)</p> <p>[E1] http://www.ncbi.nlm.nih.gov/Baxevani/HISTONES/index.html</p> <p>Histone H2B is one of the four histones, along with H2A, H3 and H4, which forms the eukaryotic nucleosome core. Using alignments of histone H2B sequences [1,2,E1], we selected a conserved region in the C-terminal part of H2B.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [KR]-E-[LIVM]-[EQ]-T-x(2)-[KR]-x-[LIVM](2)-x-[PAG]-[DE]-L-x-[KR]-H-A-[LIVM]-[STA]-E-G Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update November 1995 / Pattern and text revised.</p> <p>References [1] Wells D.E., Brown D. Nucleic Acids Res. 19:2173-2188(1991).</p> <p>[2] Thatcher T H., Gorovsky M A. Nucleic Acids Res 22:174-179(1994).</p> <p>[E1] http://www.ncbi.nlm.nih.gov/Baxevani/HISTONES/index.html</p>
--	--	---

912

HMA	PDOC00804	Heavy-metal-associated domain	<p>A conserved domain of about 30 amino acid residues has been found [1] in a number of proteins that transport or detoxify heavy metals. This domain contains two conserved cysteines that could be involved in the binding of these metals. The domain has been termed Heavy-Metal-Associated (HMA). It has been found in:</p> <ul style="list-style-type: none"> - A variety of cation transport ATPases (E1-E2 ATPases) (see <PDOC00139>). The human copper ATPases ATP7A and ATP7B which are respectively involved in Menke's and Wilson's diseases. ATP7A and ATP7B both contain 6 tandem copies of the HMA domain. The copper ATPases CCC2 from budding yeast, copA from <i>Enterococcus faecalis</i> and synA from <i>Synechococcus</i> contain one copy of the HMA domain. The cadmium ATPases cadA from <i>Bacillus firmus</i> and from plasmid p1258 from <i>Staphylococcus aureus</i> also contain a single HMA domain, while a chromosomal <i>Staphylococcus aureus</i> cadA contains two copies. Other, less characterized ATPases that contain the HMA domain are fixI from <i>Rhizobium meliloti</i>, pacS from <i>Synechococcus</i> strain PCC 7942), <i>Mycobacterium leprae</i> ctpA and ctpB and <i>Escherichia coli</i> hypothetical protein yhhO. In all these ATPases the HMA domain(s) are located in the N-terminal section. - Mercuric reductase (EC 1.16.1.1) (gene merA) which is generally encoded by plasmids carried by mercury-resistant Gram-negative bacteria. Mercuric reductase is a class-1 pyridine nucleotide-disulphide oxidoreductase (see <PDOC00073>). There is generally one HMA domain (with the exception of a chromosomal merA from <i>Bacillus</i> strain RC607 which has two) in the N-terminal part of merA. - Mercuric transport protein periplasmic component (gene merP), also encoded by plasmids carried by mercury-resistant Gram-negative bacteria. It seems to be a mercury scavenger that specifically binds to one Hg(2+) ion and which passes it to the mercuric reductase via the merT protein. The N-terminal half of merP is a HMA domain. - <i>Helicobacter pylori</i> copper-binding protein copP - Yeast protein ATX1 [2], which could act in the transport and/or partitioning of copper <p>The consensus pattern for HMA spans the complete domain.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVNS]-x(2)-[LIVMFA]-x-C-x-[STAGCDNH]-C-x(3)-[LIVFG]-x(3)-[LIV]-x(9,11)-[IVA]-x-[LVFYS] [The two C's probably bind metals] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 6. Last update December 1999 / Pattern and text revised References [1] Bull P.C., Cox D.W. Trends Genet. 10:246-252(1994) [2] Lin S.-J., Culotta V.L. Proc. Natl. Acad. Sci. U.S.A. 92:3784-3788(1995)</p>
HMG-CoA_red	PDOC00064	Hydroxymethylglutaryl-coenzyme A reductase signatures and profile	<p>Hydroxymethylglutaryl-coenzyme A reductase (EC 1.1.1.34) (HMG-CoA reductase) [1,2] catalyzes the NADP-dependent synthesis of mevalonate from 3-hydroxy-3-methylglutaryl-CoA. In vertebrates, HMG-CoA reductase is the rate-limiting enzyme in cholesterol biosynthesis. In plants, mevalonate is the precursor of all isoprenoid compounds.</p> <p>HMG-CoA reductase is a membrane bound enzyme. Structurally, it consists of 3</p>

913

			<p>domains. An N-terminal region that contains a variable number of transmembrane segments (7 in mammals, insects and fungi; 2 in plants), a linker region and a C-terminal catalytic domain of approximately 400 amino-acid residues</p> <p>In archebacteria [3] HMG-CoA reductase which is involved in the biosynthesis of the isoprenoids side chains of lipids, seems to be cytoplasmic and lack the N-terminal hydrophobic domain.</p> <p>Some bacteria, such as <i>Pseudomonas mevalonii</i>, can use mevalonate as the sole carbon source. These bacteria use an NAD-dependent HMG-CoA reductase (EC 1.1.1.88) to deacetylate mevalonate into 3-hydroxy-3-methylglutaryl-CoA [3]. The <i>Pseudomonas</i> enzyme is structurally related to the catalytic domain of NADP-dependent HMG-CoA reductases.</p> <p>We selected three conserved regions as signature patterns for HMG-CoA reductases. The first is located in the center of the catalytic domain, the second is a glycine-rich region located in the C-terminal section of the same catalytic domain and the third is also located in the C-terminal section and contains an histidine residue that seems [4] to be implicated in the catalytic mechanism as a general base.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [RKH]-x(6)-D-x-M-G-x-N-x-[LIVMA] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 4.</p> <p>Consensus pattern [LIVM]-G-x-[LIVM]-G-G-[AG]-T Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 5.</p> <p>Consensus pattern A-[LIVM]-x-[STAN]-x(2)-[LI]-x-[KRNQ]-[GSA]-H-[LM]-x-[FYLH] [H is an active site residue] Sequences known to belong to this class detected by the pattern ALL, except for archaeobacterial HMG-CoA reductases. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Sequences known to belong to this class detected by the profile ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so.</p> <p>Last update November 1997 / Patterns and text revised; profile added.</p> <p>References [1] Caelles C., Ferrer A., Balcells L., Hegardt F.G., Boronat A. Plant Mol. Biol. 13:627-638(1989).</p> <p>[2] Basson M.E., Thorsness M., Finer-Moore J., Stroud R.M., Rine J. Mol. Cell. Biol. 8:3797-3808(1988).</p> <p>[3] Lam W.L., Doolittle W.F. J. Biol. Chem. 267:5829-5834(1992).</p> <p>[4] Beach M.J., Rodwell V.W. J. Bacteriol. 171:2994-3001(1989).</p> <p>[5] Darnay B.G., Wang Y., Rodwell V.W. J. Biol. Chem. 267:15064-15070(1992).</p>
HMGL-like	PDOC00813 PDOC00643	Hydroxymethylglutaryl-coenzyme A	3-hydroxy-3-methylglutaryl-coenzyme A lyase (HMG-CoA lyase or HL) (EC 4.1.3.4) catalyzes the transformation of HMG-CoA into acetyl-CoA and acetoacetate. In

914

		<p>lyase active site; Alpha-isopropylmalate and homocitrate synthases signatures</p>	<p>vertebrates it is a mitochondrial enzyme which is involved in ketogenesis and in leucine catabolism [1]. In some bacteria, such as <i>Pseudomonas mevalonii</i>, it is involved in mevalonate catabolism (gene <i>mvaB</i>). A cysteine has been shown [2], in <i>mvaB</i>, to be required for the activity of the enzyme. The region around this residue is perfectly conserved and is used as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern S-V-A-G-L-G-G-C-P-Y [C is the active site residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE Last update November 1995 / First entry. References [1] Mitchell G.A., Robert M.-F., Hruz P.W., Wang S., Fontaine G., Behnke C.E., Mende-Mueller L.M., Schappert K., Lee C., Gibson K.M., Miziorko H.M. <i>J. Biol. Chem.</i> 268 4376-4381(1993). [2] Hruz P.W., Narasimhan C., Miziorko H.M. <i>Biochemistry</i> 31.6842-6847(1992).</p> <p>The following enzymes have been shown [1] to be functionally as well as evolutionary related:</p> <ul style="list-style-type: none"> - Alpha-isopropylmalate synthase (EC 4.1.3.12) which catalyzes the first step in the biosynthesis of leucine, the condensation of acetyl-CoA and alpha-ketoisovalerate to form 2-isopropylmalate synthase. - Homocitrate synthase (EC 4.1.3.21) (gene <i>nifV</i>) which is involved in the biosynthesis of the iron-molybdenum cofactor of nitrogenase and catalyzes the condensation of acetyl-CoA and alpha-ketoglutarate into homocitrate. - Soybean late nodulin 56. - <i>Methanococcus jannaschii</i> hypothetical proteins MJ0503, MJ1195 and MJ1392. <p>We have selected two conserved regions as signature patterns for these enzymes. The first region is located in the N-terminal section while the second region is located in the central section and contains two conserved histidine residues which could be implicated in the catalytic mechanism</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern L-R-[DE]-G-x-Q-x(10)-K Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [LIVMFV]-x(2)-H-x-H-[DN]-D-x-G-x-[GAS]-x-[GASLI] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE Last update November 1997 / Patterns and text revised. References [1] Wang S.-Z., Dean D.R., Chen J.-S., Johnson J.L. <i>J. Bacteriol.</i> 173.3041-3046(1991).</p>
hormone5	PDOC00237	Neurohypophyseal hormones signature	<p>Oxytocin (or ocytocin) and vasopressin [1] are small (nine amino acid residues), structurally and functionally related neurohypophyseal peptide hormones. Oxytocin causes contraction of the smooth muscle of the uterus and of the mammary gland while vasopressin has a direct antidiuretic action on the kidney and also causes vasoconstriction of the peripheral vessels. Like the majority of active peptides, both hormones are synthesized as larger protein precursors that are enzymatically converted to their mature forms. Peptides belonging to this family are also found in birds, fish, reptiles and amphibians (mesotocin, isotocin, valitocin, glutitocin, aspartocin, vasotocin, seritocin, asvatocin, phasvatocin), in worms (annetocin), octopi</p>

915

			<p>(cephalotocin), locust (locupressin or neuropeptide F1/F2) and in molluscs (conopressins G and S) [2].</p> <p>The pattern developed to detect this category of peptides spans their entire sequence and includes four invariant amino acid residues.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-[LIFY](2)-x-N-[CS]-P-x-G [The two C's are linked by a disulfide bond].</p> <p>Sequences known to belong to this class detected by the pattern ALL</p> <p>Other sequence(s) detected in SWISS-PROT NONE</p> <p>Last update November 1995 / Pattern and text revised</p> <p>References [1] Acher R , Chauvet J Biochimie 70:1197-1207(1988).</p> <p>[2] Chauvet J., Michel G., Ouedraogo Y., Chou J., Chait B T., Acher R. Int. J. Pept. Protein Res. 45:482-487(1995)</p>
HPPK	PDOC00631	7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase signature	<p>All organisms require reduced folate cofactors for the synthesis of a variety of metabolites. Most microorganisms must synthesize folate de novo because they lack the active transport system of higher vertebrate cells which allows these organisms to use dietary folates. Enzymes involved in folate biosynthesis are therefore targets for a variety of antimicrobial agents such as trimethoprim or sulfonamides</p> <p>7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (EC 2.7.6.3) (HPPK) catalyzes the attachment of pyrophosphate to 6-hydroxymethyl-7,8-dihydropterin to form 6-hydroxymethyl-7,8-dihydropteridine pyrophosphate. This is the first step in a three-step pathway leading to 7,8-dihydrofolate.</p> <p>Bacterial HPPK (gene folK or sulD) [1] is a protein of 160 to 270 amino acids. In the lower eukaryote <i>Pneumocystis carinii</i>, HPPK is the central domain of a multifunctional folate synthesis enzyme (gene fas) [2].</p> <p>As a signature for HPPK, we selected a conserved region located in the central section of these enzymes</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [KRHD]-x-[GA]-[PSAE]-R-x(2)-D-[LIV]-D-[LIVM](2)</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE</p> <p>Last update July 1999 / Pattern and text revised.</p> <p>References [1] Talarico T.L., Ray P.H., Dev I.K., Merrill B.M., Dallas W.S. J. Bacteriol. 174:5971-5977(1992).</p> <p>[2] Volpes F., Dyer M., Scaife J.G., Darby G., Stammers D.K., Delves C.J. Gene 112:213-218(1992).</p>
Hydrolase		haloacid dehalogenase-like hydrolase	<p>Accession number: PF00702</p> <p>Definition: haloacid dehalogenase-like hydrolase</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_566 (release 2.1)</p> <p>Gathering cutoffs: 7.7</p> <p>Trusted cutoffs: 7.10 7.10</p> <p>Noise cutoffs: 2.90 2.90</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p>

			<p>Reference Medline: 96355356</p> <p>Reference Title: Crystal structure of L-2-haloacid dehalogenase from</p> <p>Reference Title: Pseudomonas sp. YL. An alpha/beta hydrolase structure that</p> <p>Reference Title: is different from the alpha/beta hydrolase fold.</p> <p>Reference Author: Hisano T, Hata Y, Fujii T, Liu JQ, Kurihara T, Esaki N, Soda K;</p> <p>Reference Author: Soda K;</p> <p>Reference Location: J Biol Chem 1996;271:20322-20330.</p> <p>Database Reference: SCOP; 1jud; sf. [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO, IPR001454;</p> <p>Database Reference: PDB; 1jud ; 4; 197;</p> <p>Database Reference: PDB, 1zrm ; 4; 197;</p> <p>Database Reference: PDB, 1zrn ; 4; 197;</p> <p>Database Reference: PDB, 1aq6 A; 2; 193;</p> <p>Database Reference: PDB, 1aq6 B; 2; 193;</p> <p>Database Reference: PDB, 1qq5 A; 2; 193;</p> <p>Database Reference: PDB, 1qq5 B; 2; 193;</p> <p>Database Reference: PDB; 1qq6 A; 2; 193;</p> <p>Database Reference: PDB; 1qq6 B; 2; 193;</p> <p>Database Reference: PDB; 1qq7 A; 2; 193;</p> <p>Database Reference: PDB; 1qq7 B; 2; 193;</p> <p>Database Reference: PDB; 1cqz A; 4; 19;</p> <p>Database Reference: PDB, 1cr6 A; 4; 19;</p> <p>Database Reference: PDB, 1cqz B; 4; 206;</p> <p>Database Reference: PDB, 1cr6 B; 4; 206;</p> <p>Database Reference: PDB, 1cqz A; 48; 206;</p> <p>Database Reference: PDB; 1cr6 A, 48; 206;</p> <p>Database reference: PFAMB; PB000701;</p> <p>Database reference: PFAMB; PB001048;</p> <p>Database reference: PFAMB; PB019234;</p> <p>Database reference: PFAMB; PB032787;</p> <p>Database reference: PFAMB; PB040985;</p> <p>Database reference: PFAMB; PB041061;</p> <p>Database reference: PFAMB; PB041182;</p> <p>Database reference: PFAMB; PB041477;</p> <p>Database reference: PFAMB; PB041535;</p> <p>Database reference: PFAMB; PB041628;</p> <p>Database reference: PFAMB; PB041677;</p> <p>Comment: This family are structurally different from the alpha/</p> <p>Comment: beta hydrolase family (abhydrolase).</p> <p>Comment: This family includes L-2-haloacid dehalogenase, epoxide</p> <p>Comment: hydrolases and phosphatases.</p> <p>Comment: The structure of the family consists of two domains. One</p> <p>Comment: is an inserted four helix bundle, which is the least well</p> <p>Comment: conserved region of the alignment. between residues 16</p> <p>Comment: and</p> <p>Comment: 96 of Swiss:P24069 The rest of the fold is composed of</p> <p>Comment: the</p> <p>Comment: core alpha/beta domain.</p> <p>Number of members: 134</p>
HypB_UreG		HypB/UreG nucleotide-binding domain	<p>Accession number: PF01495</p> <p>Definition: HypB/UreG nucleotide-binding domain</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_428 (release 4.0)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 197.70 197.70</p> <p>Noise cutoffs: -40.00 -40.00</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmbuild --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97285753</p> <p>Reference Title: The HypB protein from Bradyrhizobium japonicum can</p> <p>Reference Title: store</p> <p>Reference Title: nickel and is required for the nickel-dependent</p> <p>Reference Title: transcriptional regulation of hydrogenase.</p> <p>Reference Author: Olson JW, Fu C, Maier RJ;</p> <p>Reference Location: Mol Microbiol 1997;24:119-128</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97352660</p> <p>Reference Title: Characterization of UreG, identification of a</p> <p>Reference Title: UreD-UreF-UreG complex, and evidencesuggesting that</p> <p>Reference Title: a</p>

			<p>Reference Title: nucleotide-binding site in UreG is required for in vivo metallocenter assembly of Klebsiella aerogenes urease.</p> <p>Reference Title: metallocenter assembly of Klebsiella aerogenes urease.</p> <p>Reference Author: Moncrief MB, Hausinger RP;</p> <p>Reference Location: J Bacteriol 1997;179:4081-4086.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 93139028</p> <p>Reference Title: The product of the hypB gene, which is required for nickel incorporation into hydrogenases, is a novel guanine nucleotide-binding protein.</p> <p>Reference Title: nucleotide-binding protein.</p> <p>Reference Author: Maier T, Jacobi A, Sauter M, Bock A;</p> <p>Reference Location: J Bacteriol 1993;175:630-635.</p> <p>Reference Number: [4]</p> <p>Reference Medline: 92325016</p> <p>Reference Title: Klebsiella aerogenes urease gene cluster. sequence of ureD</p> <p>Reference Title: and demonstration that four accessory genes (ureD, ureE, ureF, and ureG) are involved in nickel metallocenter biosynthesis.</p> <p>Reference Title: ureF, and ureG) are involved in nickel metallocenter biosynthesis.</p> <p>Reference Author: Lee MH, Mulrooney SB, Renner MJ, Markowicz Y, Hausinger RP;</p> <p>Reference Location: J Bacteriol 1992;174:4324-4330.</p> <p>Database Reference: INTERPRO, IPR002894:</p> <p>Comment: This domain is found in HypB a hydrogenase expression / formation</p> <p>Comment: protein. and UreG a urease accessory protein. Both these proteins contain</p> <p>Comment: a P-loop nucleotide binding motif [2.3]. HypB has GTPase activity</p> <p>Comment: and is a guanine nucleotide binding protein [3]. It is not known</p> <p>Comment: whether UreG binds GTP or some other nucleotide. Both enzymes are involved</p> <p>Comment: in nickel binding. HypB can store nickel and is required for nickel</p> <p>Comment: dependent hydrogenase expression [1]. UreG is required for functional</p> <p>Comment: incorporation of the urease nickel metallocenter.[4] GTP hydrolysis may</p> <p>Comment: required by these proteins for nickel incorporation into other nickel</p> <p>Comment: proteins [1]</p> <p>Number of members: 41</p>
IBB	Importin beta binding domain		<p>Accession number: PF01749</p> <p>Definition: Importin beta binding domain</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members Pfam-B_544 (release 4.2)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 67 30 67.30</p> <p>Noise cutoffs: -15.90 -15.90</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmbuild --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98359119</p> <p>Reference Title: Crystallographic analysis of the recognition of a nuclear localization signal by the nuclear import factor karyopherin alpha.</p> <p>Reference Title: karyopherin alpha.</p> <p>Reference Author: Conti E, Uy M, Leighton L, Blobel G, Kuriyan J;</p> <p>Reference Location: Cell 1998;94:193-204.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 98275030</p> <p>Reference Title: Importins and exportins: how to get in and out of the nucleus [published erratum appears in Trends Biochem Sci</p> <p>Reference Title: 1998 Jul;23(7):235]</p> <p>Reference Author: Weis K;</p> <p>Reference Location: Trends Biochem Sci 1998;23:185-189.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 98250643</p> <p>Reference Title: Transport into and out of the cell nucleus.</p> <p>Reference Author: Gorlich D;</p> <p>Reference Location: EMBO J 1998;17:2721-2727.</p>

918

			<p>Reference Number: [4] Reference Medline: 96270582 Reference Title: The binding site of karyopherin alpha for karyopherin beta overlaps with a nuclear localization sequence. Reference Author: Moroianu J, Blobel G, Radu A; Reference Location: Proc Natl Acad Sci U S A 1996;93:6572-6576. Reference Number: [5] Reference Medline: 96203101 Reference Title: A 41 amino acid motif in importin-alpha confers binding to importin- beta and hence transit into the nucleus. Reference Author: Gorlich D, Henklein P, Laskey RA, Hartmann E; Reference Location: EMBO J 1996;15:1810-1817. Database Reference: SCOP, 1bk5; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO, IPR002652; Database Reference: PDB; 1ejl I; 72; 99; Database Reference: PDB; 1ejy I; 72; 99; Database Reference: PDB; 1lal A; 44; 99; Database Reference: PDB; 1qqr B; 28; 51; Database Reference: PDB; 1qgk B; 11; 54; Database Reference: PDB; 1ee5 A; 90; 110; Database Reference: PDB; 1bk5 A; 89; 110; Database Reference: PDB; 1bk5 B; 89; 110; Database Reference: PDB; 1bk6 A; 89; 110; Database Reference: PDB; 1bk6 B; 89; 110; Database Reference: PDB; 1ee4 A; 87; 110; Database Reference: PDB; 1ee4 B; 87; 110; Comment: This family consists of the importin alpha (karyopherin alpha), Comment: importin beta (karyopherin beta) binding domain. The domain mediates Comment: formation of the importin alpha beta complex: required for classical Comment: NLS import of proteins into the nucleus, through the nuclear pore Comment: complex and across the nuclear envelope. Comment: Also in the alignment is the NLS of importin alpha which overlaps Comment: with the IBB domain [4] Number of members: 38</p>
IF-2B		Initiation factor 2 subunit family	<p>Accession number: PF01008 Definition: Initiation factor 2 subunit family Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1302 (release 3.0) Gathering cutoffs: -135 -135 Trusted cutoffs: -82.40 -82.40 Noise cutoffs: -157.30 -157.30 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98188271 Reference Title: Archaeal translation initiation revisited: the initiation factor 2 and eukaryotic initiation factor 2B Reference Title: alpha-beta-delta subunit families. Reference Author: Kyrpides NC, Woese CR; Reference Location: Proc Natl Acad Sci U S A 1998;95:3726-3730. Database Reference: INTERPRO; IPR000649; Comment: This family includes initiation factor 2B alpha, beta and delta Comment: subunits from eukaryotes, initiation factor 2B subunits 1 and 2 Comment: from archaeobacteria and some proteins of unknown function from Comment: prokaryotes. Initiation factor 2 binds to Met-tRNA, GTP and the Comment: small ribosomal subunit. Number of members: 33</p>
IF3	PDOC00723	Initiation factor 3 signature	<p>Initiation factor 3 (IF-3) (gene infC) [1] is one of the three factors required for the initiation of protein biosynthesis in bacteria. IF-3 is thought to function as a fidelity factor during the assembly of the ternary initiation complex which consist of the 30S ribosomal subunit, the initiator tRNA and the messenger RNA. IF-3 binds to the 30S ribosomal subunit; it is a</p>

			<p>basic protein of 141 to 212 residues.</p> <p>The chloroplast initiation factor IF-3(chl) is a protein that enhances the poly(A,U,G)-dependent binding of the initiator tRNA to chloroplast ribosomal 30s subunits. In its mature form it is a protein of about 400 residues whose central section is evolutionary related to the sequence of bacterial IF-3 [2].</p> <p>As a signature pattern we selected a highly conserved region located in the central section of bacterial IF-3 and of IF-3(chl)</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [KR]-[LIVM]{2}-[DN]-[FY]-[GSN]-[KR]-[LIVMFYS]-x-[FY]-[DEQTH]-x(2)-[KRQ]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update July 1999 / Pattern and text revised</p> <p>References [1] Liveris D., Schwartz J.J., Geertman R., Schwartz I. FEMS Microbiol. Lett. 112:211-216(1993).</p> <p>[2] Lin Q., Ma L., Burkhart W., Spremulli L.L. J. Biol. Chem. 269:9436-9444(1994).</p>
IF4E	PDOC00641	Eukaryotic initiation factor 4E signature	<p>Eukaryotic translation initiation factor 4E (eIF-4E) [1] is a protein that binds to the cap structure of eukaryotic cellular mRNAs. eIF-4E recognizes and binds the 7-methylguanosine-containing (m7Gppp) cap during an early step in the initiation of protein synthesis and facilitates ribosome binding to a mRNA by inducing the unwinding of its secondary structures.</p> <p>eIF-4E is a conserved protein of about 25 Kd. Site directed mutagenesis experiments have shown [2] that a tryptophan in the central part of the sequence of human eIF-4E seems to be implicated in cap-binding. The signature pattern for eIF-4E includes this tryptophan.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [DE]-[IFY]-x(2)-F-[KR]-x(2)-[LIVM]-x-P-x-W-E-[DVA]-x(5)-G-G-[KR]-W [The first W seems to be involved in cap-binding]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE</p> <p>Last update July 1999 / Pattern and text revised.</p> <p>References [1] Thach R.E. Cell 68:177-180(1992).</p> <p>[2] Ueda H., Iyo H., Doi M., Inoue M., Ishida T., Morioka H., Tanaka T., Nishikawa S., Uesugi S. FEBS Lett. 280:207-210(1991).</p>
IF5_eIF4_eIF2		eIF4-gamma/eIF5/eIF2-epsilon	<p>Accession number: PF02020</p> <p>Definition: eIF4-gamma/eIF5/eIF2-epsilon</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: [1]</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 26 10 26 10</p> <p>Noise cutoffs: -21.50 -21 50</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96060092</p> <p>Reference Title: Multidomain organization of eukaryotic guanine</p>

			<p>nucleotide</p> <p>Reference Title exchange translation initiation factor eIF-2B subunits</p> <p>Reference Title: revealed by analysis of conserved sequence motifs.</p> <p>Reference Author: Koonin EV;</p> <p>Reference Location Protein Sci 1995;4:1608-1617</p> <p>Comment: This domain of unknown function is found at the C-terminus</p> <p>Comment: of several transcription initiation factors [1].</p> <p>Number of members: 31</p>
ig	PDOC00262	Immunoglobulins and major histocompatibility complex proteins signature	<p>The basic structure of immunoglobulin (Ig) [1] molecules is a tetramer of two light chains and two heavy chains linked by disulfide bonds. There are two types of light chains: kappa and lambda, each composed of a constant domain (CL) and a variable domain (VL). There are five types of heavy chains: alpha, delta, epsilon, gamma and mu, all consisting of a variable domain (VH) and three (in alpha, delta and gamma) or four (in epsilon and mu) constant domains (CH1 to CH4).</p> <p>The major histocompatibility complex (MHC) molecules are made of two chains.</p> <p>In class I [2] the alpha chain is composed of three extracellular domains, a transmembrane region and a cytoplasmic tail. The beta chain (beta-2-microglobulin) is composed of a single extracellular domain. In class II [3], both the alpha and the beta chains are composed of two extracellular domains, a transmembrane region and a cytoplasmic tail.</p> <p>It is known [4,5] that the Ig constant chain domains and a single extracellular domain in each type of MHC chains are related. These homologous domains are approximately one hundred amino acids long and include a conserved intradomain disulfide bond. We developed a small pattern around the C-terminal cysteine involved in this disulfide bond which can be used to detect these category of Ig related proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [FY]-x-C-x-[VA]-x-H-Sequences known to belong to this class detected by the pattern: Ig heavy chains type Alpha C region : All, in CH2 and CH3. Ig heavy chains type Delta C region : All, in CH3. Ig heavy chains type Epsilon C region: All, in CH1, CH3 and CH4. Ig heavy chains type Gamma C region : All, in CH3 and also CH1 in some cases. Ig heavy chains type Mu C region : All, in CH2, CH3 and CH4. Ig light chains type Kappa C region : In all CL except rabbit and Xenopus. Ig light chains type Lambda C region : In all CL except rabbit. MHC class I alpha chains : All, in alpha-3 domains, including in the cytomegalovirus MHC-1 homologous protein [6]. Beta-2-microglobulin : All. MHC class II alpha chains: All, in alpha-2 domains. MHC class II beta chains. All, in beta-2 domains.</p> <p>Other sequence(s) detected in SWISS-PROT 71</p> <p>Last update</p> <p>May 1991 / Text revised.</p> <p>References</p> <p>[1]</p> <p>Gough N.</p> <p>Trends Biochem. Sci. 6:203-205(1981)</p> <p>[2]</p> <p>Klein J , Figueroa F.</p> <p>Immunol. Today 7 41-44(1986)</p> <p>[3]</p> <p>Figueroa F , Klein J.</p> <p>Immunol. Today 7:78-81(1986).</p> <p>[4]</p> <p>Orr H.T , Lancet D., Robb R.J., Lopez de Castro J.A., Strominger J.L.</p> <p>Nature 282:266-270(1979).</p> <p>[5]</p> <p>Cushley W., Owen M.J.</p> <p>Immunol. Today 4:88-92(1983).</p> <p>[6]</p>

			Beck S., Barrel B.G. Nature 331:269-272(1988).
IMPDH_C	PDOC00391	IMP dehydrogenase / GMP reductase signature	<p>IMP dehydrogenase (EC 1.1.1.205) (IMPDH) catalyzes the rate-limiting reaction of de novo GTP biosynthesis, the NAD-dependent reduction of IMP into XMP [1]. Inhibition of IMP dehydrogenase activity results in the cessation of DNA synthesis. As IMP dehydrogenase is associated with cell proliferation, it is a possible target for cancer chemotherapy. Mammalian and bacterial IMPDHs are tetramers of identical chains. There are two IMP dehydrogenase isozymes in humans [2].</p> <p>GMP reductase (EC 1.6.6.8) catalyzes the irreversible and NADPH-dependent reductive deamination of GMP into IMP [3]. It converts nucleobase, nucleoside and nucleotide derivatives of G to A nucleotides, and maintains intracellular balance of A and G nucleotides.</p> <p>IMP dehydrogenase and GMP reductase share many regions of sequence similarity. One of these regions is centered on a cysteine residue thought [3] to be involved in binding IMP. We have used this region as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVM]-[RK]-[LIVM]-G-[LIVM]-G-x-G-S-[LIVM]-C-x-T [C is the putative IMP-binding residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE Last update May 1991 / First entry References [1] Collart F.R., Huberman E. J. Biol. Chem. 263.15769-15772(1988).</p> <p>[2] Natsumeda Y., Ohno S., Kawasaki H., Konno Y., Weber G., Suzuki K J. Biol. Chem. 265.5292-5295(1990).</p> <p>[3] Andrews S.C., Guest J.R. Biochem. J. 255.35-43(1988)</p>
Inos-1-P_synth		Myo-inositol-1-phosphate synthase	<p>Accession number: PF01658 Definition: Myo-inositol-1-phosphate synthase Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_959 (release 4 1) Gathering cutoffs: 25 25 Trusted cutoffs: 86 80 86.80 Noise cutoffs: -219.00 -219.00 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 95066381 Reference Title: Comparison of INO1 gene sequences and products in <i>Candida albicans</i> and <i>Saccharomyces cerevisiae</i>. Reference Author: Klig LS, Zobel PA, Devry CG, Losberger C; Reference Location: Yeast 1994;10:789-800. Database Reference: INTERPRO; IPR002587; Comment: This is a family of myo-inositol-1-phosphate synthases. Comment: Inositol-1-phosphate catalyses the conversion of glucose-6-phosphate to inositol-1-phosphate, which is then dephosphorylated Comment: to inositol [1]. Inositol phosphates play an important role in signal transduction Number of members: 27</p>

IPP_isomerase		Isopentenyl-diphosphate delta-isomerase	<p>Accession number. PF01772</p> <p>Definition: Isopentenyl-diphosphate delta-isomerase</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B 1099 (release 4.2)</p> <p>Gathering cutoffs: -88 -88</p> <p>Trusted cutoffs: -66.70 -66.70</p> <p>Noise cutoffs: -106.90 -106.90</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98409684</p> <p>Reference Title: Differential expression of two isopentenyl pyrophosphate isomerases and enhanced carotenoid accumulation in a unicellular chlorophyte</p> <p>Reference Author: Sun Z, Cunningham FX Jr, Gantt E;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1998;95:11482-11488.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97373600</p> <p>Reference Title: Cloning and subcellular localization of hamster and rat isopentenyl diphosphate dimethylallyl diphosphate isomerase. A PTS1 motif targets the enzyme to peroxisomes.</p> <p>Reference Author: Paton VG, Shackelford JE, Krisans SK;</p> <p>Reference Location: J Biol Chem 1997;272:18945-18950.</p> <p>Database Reference: INTERPRO; IPR002667,</p> <p>Comment: Isopentenyl-diphosphate delta-isomerase or IPP isomerase EC:5.3.3.2</p> <p>Comment: catalyses the interconversion of isopentenyl diphosphate and dimethylallyl diphosphate. Dimethylallyl phosphate is the initial substrate</p> <p>Comment: for the biosynthesis of carotenoids and other long chain isoprenoids [1].</p> <p>Number of members: 24</p>
K-box	PDOC00302	MADS-box domain signature and profile	<p>A number of transcription factors contain a conserved domain of 56 amino-acid residues, sometimes known as the MADS-box domain [E1]. They are listed below:</p> <ul style="list-style-type: none"> - Serum response factor (SRF) [1], a mammalian transcription factor that binds to the Serum Response Element (SRE). This is a short sequence of dyad symmetry located 300 bp to the 5' end of the transcription initiation site of genes such as c-fos. - Mammalian myocyte-specific enhancer factors 2A to 2D (MEF2A to MEF2D). <p>These proteins are transcription factor which binds specifically to the MEF2 element present in the regulatory regions of many muscle-specific genes.</p> <ul style="list-style-type: none"> - Drosophila myocyte-specific enhancer factor 2 (MEF2). - Yeast GRM/PRTF protein (gene MCM1) [2], a transcriptional regulator of mating-type-specific genes - Yeast arginine metabolism regulation protein I (gene ARG81 or ARG80). - Yeast transcription factor RLM1. - Yeast transcription factor SMP1 - Arabidopsis thaliana agamous protein (AG) [3], a probable transcription factor involved in regulating genes that determines stamen and carpel development in wild-type flowers. Mutations in the AG gene result in the replacement of the stamens by petals and the carpels by a new flower. - Arabidopsis thaliana homeotic proteins Apetala1 (AP1), Apetala3 (AP3) and Pistillata (PI) which act locally to specify the identity of the floral meristem and to determine sepal and petal development [4]. - Antirrhinum majus and tobacco homeotic protein deficiens (DEFA) and globosa (GLO) [5]. Both proteins are transcription factors involved in the genetic control of flower development. Mutations in DEFA or GLO cause the transformation of petals into sepals and of stamens into carpels. - Arabidopsis thaliana putative transcription factors AGL1 to AGL6 [6] - Antirrhinum majus morphogenetic protein DEF H33 (squamosa). <p>In SRF, the conserved domain has been shown [1] to be involved in DNA-binding</p>

923

			<p>and dimerization. We have derived a pattern that spans the complete length of the domain. The profile also spans the length of the MADS-box.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern R-x-[RK]-x(5)-I-x-[DNGSK]-x(3)-[KR]-x(2)-T-[FY]-x-[RK](3)-x(2)-[LIVM]-x-K(2)-A-x-E-[LIVM]-[STA]-x-L-x(4)-[LIVM]-x-[LIVM](3)-x(6)-[LIVMF]-x(2)-[FY]</p> <p>Sequences known to belong to this class detected by the pattern ALL</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Sequences known to belong to this class detected by the profile ALL</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note this documentation entry is linked to both signature patterns and a profile. As the profile is much more sensitive than the patterns, you should use it if you have access to the necessary software tools to do so.</p> <p>Last update July 1999 / Pattern and text revised.</p> <p>References</p> <p>[1] Norman C. Runswick M., Pollock R., Treisman R. Cell 55:989-1003(1988).</p> <p>[2] Passmore S., Maine G.T., Elble R., Christ C., Tye B.-K. J. Mol. Biol. 204:593-606(1988).</p> <p>[3] Yanofsky M., Ma H., Bowman J., Drews G., Feldmann K.A., Meyerowitz E.M. Nature 346:35-39(1990)</p> <p>[4] Goto K., Meyerowitz E.M. Genes Dev. 8:1548-1560(1994)</p> <p>[5] Troebner W., Ramirez L., Motte P., Hue I., Huijser P., Loennig W.-E., Saedler H., Sommer H., Schwartz-Sommer Z. EMBO J. 11:4693-4704(1992)</p> <p>[6] Ma H., Yanofsky M.F., Meyerowitz E.M. Genes Dev. 5:484-495(1991)</p> <p>[E1] http://transfac.gbf-braunschweig.de/cgi-bin/qt/getEntry.pl?C0014</p>
Keratin_B2		Keratin, high sulfur B2 protein	<p>Accession number PF01500</p> <p>Definition: Keratin, high sulfur B2 protein</p> <p>Author: Bateman A</p> <p>Alignment method of seed Clustalw</p> <p>Source of seed members: Pfam-B_706 (release 4.0)</p> <p>Gathering cutoffs: -17 -17</p> <p>Trusted cutoffs: -1.50 -1.50</p> <p>Noise cutoffs: -46.00 18.50</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98201605</p> <p>Reference Title: Structure and hair follicle-specific expression of genes encoding the rat high sulfur protein B2 family.</p> <p>Reference Title: encoding the rat high sulfur protein B2 family.</p> <p>Reference Author: Mitsui S, Ohuchi A, Adachi-Yamada T, Hotta M, Tsuboi R,</p> <p>Reference Author: Ogawa H;</p> <p>Reference Location: Gene 1998;208:123-129.</p> <p>Database Reference INTERPRO: IPR002494,</p> <p>Comment: High sulfur proteins are cysteine-rich proteins synthesized during the differentiation of hair matrix cells, and form hair fibers in association with hair keratin intermediate filaments [1].</p> <p>Comment: This family has been divided up into four regions, with the</p>

924

			<p>second</p> <p>Comment: region containing 8 copies of a short repeat [1] This family is</p> <p>Comment: also known as B2 or KAP1.</p> <p>Number of members: 17</p>
ketoacyl-synt	PDOC00529	Beta-ketoacyl synthases active site	<p>Beta-ketoacyl-ACP synthase (EC 2.3.1.41) (KAS) [1] is the enzyme that catalyzes the condensation of malonyl-ACP with the growing fatty acid chain. It is found as a component of the following enzymatic systems</p> <ul style="list-style-type: none"> - Fatty acid synthetase (FAS), which catalyzes the formation of long-chain fatty acids from acetyl-CoA, malonyl-CoA and NADPH. Bacterial and plant chloroplast FAS are composed of eight separate subunits which correspond to different enzymatic activities; beta-ketoacyl synthase is one of these polypeptides. Fungal FAS consists of two multifunctional proteins, FAS1 and FAS2; the beta-ketoacyl synthase domain is located in the C-terminal section of FAS2. Vertebrate FAS consists of a single multifunctional chain; the beta-ketoacyl synthase domain is located in the N-terminal section [2] - The multifunctional 6-methylsalicylic acid synthase (MSAS) from <i>Penicillium patulum</i> [3]. This is a multifunctional enzyme involved in the biosynthesis of a polyketide antibiotic and which has a KAS domain in its N-terminal section. - Polyketide antibiotic synthase enzyme systems. Polyketides are secondary metabolites produced by microorganisms and plants from simple fatty acids. KAS is one of the components involved in the biosynthesis of the <i>Streptomyces</i> polyketide antibiotics granatacin [4], tetracenomycin C [5] and erythromycin. - <i>Emericella nidulans</i> multifunctional protein Wa. Wa is involved in the biosynthesis of conidial green pigment. Wa is protein of 216 Kd that contains a KAS domain. - <i>Rhizobium</i> nodulation protein nodE, which probably acts as a beta-ketoacyl synthase in the synthesis of the nodulation Nod factor fatty acyl chain. - Yeast mitochondrial protein CEM1 <p>The condensation reaction is a two step process: the acyl component of an activated acyl primer is transferred to a cysteine residue of the enzyme and is then condensed with an activated malonyl donor with the concomitant release of carbon dioxide. The sequence around the active site cysteine is well conserved and can be used as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-x(4)-[LIVMFAP]-x(2)-[AGC]-C-[STA](2)-[STAG]-x(3)-[LIVMF] [C is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for bacterial and plant beta-ketoacyl synthase III (KAS III).</p> <p>Other sequence(s) detected in SWISS-PROT 10</p> <p>Last update November 1997 / Text revised.</p> <p>References</p> <p>[1] Kauppinen S., Siggaard-Andersen M., von Wettstein-Knowles P Carlsberg Res. Commun. 53:357-370(1988).</p> <p>[2] Witkowski A., Rangan V.S., Randhawa Z.I., Amy C.M., Smith S. Eur. J. Biochem. 198:571-579(1991).</p> <p>[3] Beck J., Ripka S., Siegner A., Schiltz E., Schweizer E. Eur. J. Biochem. 192:487-498(1990).</p> <p>[4] Bibb M.J., Biro S., Motamedi H., Collins J.F., Hutchinson C.R. EMBO J. 8:2727-2736(1989).</p> <p>[5] Sherman D.H., Malpartida F., Bibb M.J., Kieser H.M., Bibb M.J., Hopwood D.A. EMBO J. 8:2717-2725(1989).</p>
KRAB		KRAB box	<p>Accession number. PF01352</p> <p>Definition: KRAB box</p>

925

			<p>Author: Bateman A</p> <p>Alignment method of seed. Manual</p> <p>Source of seed members: Bateman A</p> <p>Gathering cutoffs: 0 0</p> <p>Trusted cutoffs: 1.10 1.10</p> <p>Noise cutoffs: -5.40 -5.40</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 91319563</p> <p>Reference Title: Conserved KRAB protein domain identified upstream from the</p> <p>Reference Title: zinc finger region of Kox 8.</p> <p>Reference Author: Thiesen HJ, Bellefroid E, Revelant O, Martial JA;</p> <p>Reference Location: Nucleic Acids Res 1991;19:3996-3996</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97140325</p> <p>Reference Title: A novel member of the RING finger family, KRIP-1, associates with the KRAB-A transcriptional repressor domain</p> <p>Reference Title: of zinc finger proteins.</p> <p>Reference Author: Kim SS, Chen YM, O'Leary E, Witzgall R, Vidal M. Bonventre</p> <p>Reference Author: JV;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1996;93:15299-15304</p> <p>Reference Number: [3]</p> <p>Reference Medline: 96365472</p> <p>Reference Title: KAP-1, a novel corepressor for the highly conserved KRAB</p> <p>Reference Title: repression domain.</p> <p>Reference Author: Friedman JR, Fredericks WJ, Jensen DE, Speicher DW, Huang</p> <p>Reference Author: XP, Neilson EG, Rauscher FJ;</p> <p>Reference Location: Genes Dev 1996;10:2067-2078.</p> <p>Database Reference: INTERPRO; IPR001909;</p> <p>Database reference: PFAMB, PB036541;</p> <p>Comment: The KRAB domain (or Kruppel-associated box) is present in</p> <p>Comment: about a third of zinc finger proteins containing C2H2 fingers.</p> <p>Comment: The KRAB domain is found to be involved in protein-protein</p> <p>Comment: interactions [2.3].</p> <p>Comment: The KRAB domain is generally encoded by two exons. The</p> <p>Comment: regions coded by the two exons are known as KRAB-A and</p> <p>Comment: KRAB-B.</p> <p>Number of members: 105</p>
lectin_legB	PDOC00278	Legume lectins signatures	<p>Leguminous plants synthesize sugar-binding proteins which are called legume lectins [1,2]. These lectins are generally found in the seeds. The exact function of legume lectins is not known but they may be involved in the attachment of nitrogen-fixing bacteria to legumes and in the protection against pathogens. Legume lectins bind calcium and manganese (or other transition metals).</p> <p>Legume lectins are synthesized as precursor proteins of about 230 to 260 amino acid residues. Some legume lectins are proteolytically processed to produce two chains: beta (which corresponds to the N-terminal) and alpha (C-terminal). The lectin concanavalin A (conA) from jack bean is exceptional in that the two chains are transposed and ligated (by formation of a new peptide bond) The N-terminus of mature conA thus corresponds to that of the alpha chain and the C-terminus to the beta chain.</p> <p>We have developed two signature patterns specific to legume lectins: the first is located in the C-terminal section of the beta chain and contains a conserved aspartic acid residue important for the binding of calcium and manganese; the second one is located in the N-terminal of the alpha chain.</p> <p>Description of pattern(s) and/or profile(s)</p>

			<p>Consensus pattern [LIV]-[STAG]-V-[DEQV]-[FLI]-D-[ST] [D binds manganese and calcium] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT 21.</p> <p>Consensus pattern [LIV]-x-[EDQ]-[FYWKR]-V-x-[LIVF]-G-[LF]-[ST] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 4. Last update July 1999 / Patterns and text revised. References [1] Sharon N., Lis H FASEB J. 4:3198-320(1990).</p> <p>[2] Lis H., Sharon N. Annu. Rev. Biochem. 55:33-37(1986).</p>
ligase-CoA		CoA-ligases	<p>Accession number: PF00549 Definition: CoA-ligases Author: Bateman A Alignment method of seed: Clustalw Source of seed members: SCOP Gathering cutoffs: 25 25 Trusted cutoffs: 28 70 28.70 Noise cutoffs: 14.70 14.70 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 94193797 Reference Title: The crystal structure of succinyl-CoA synthetase from Escherichia coli at 2.5-A resolution. Reference Author: Wolodko WT, Fraser ME, James MN, Bridger WA: Reference Location: J Biol Chem 1994;269:10883-10890. Database Reference: SCOP; 1scu: sf; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR000303; Database Reference: PDB; 1cqi A; 132; 279; Database Reference: PDB; 1cqi D; 132; 279; Database Reference: PDB; 1cqi A; 132; 279; Database Reference: PDB; 1cqi D; 132; 279; Database Reference: PDB; 2scu A; 132; 279; Database Reference: PDB; 2scu D; 132; 279; Database Reference: PDB; 1scu A; 132; 279; Database Reference: PDB; 1scu D; 132; 279; Database Reference: PDB; 1cqi B; 246; 385; Database Reference: PDB; 1cqi E; 246; 385; Database Reference: PDB; 1cqi B; 246; 385; Database Reference: PDB; 1cqi E; 246; 385; Database Reference: PDB; 2scu B; 246; 385; Database Reference: PDB; 2scu E; 246; 385; Database Reference: PDB; 1scu B; 246; 388; Database Reference: PDB; 1scu E; 246; 388; Database reference: PFAMB; PB039724; Database reference: PFAMB; PB041236; Comment: - - This family includes the CoA ligases Succinyl-CoA synthetase alpha Comment: and beta chains, malate CoA ligase and ATP-citrate lyase Comment: Some members of the family utilise ATP others use GTP. Number of members: 76</p>
LIM_bind		LIM-domain binding protein	<p>Accession number: PF01803 Definition: LIM-domain binding protein Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1352 (release 4.2) Gathering cutoffs: -92 -92 Trusted cutoffs: 13.40 13.40 Noise cutoffs: -197.90 -197.90 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 97477378</p>

			<p>Reference Title: Chip, a widely expressed chromosomal protein required for segmentation and activity of a remote wing margin enhancer</p> <p>Reference Title: in Drosophila.</p> <p>Reference Author: Morcillo P, Rosen C, Baylies MK, Dorsett D;</p> <p>Reference Location: Genes Dev 1997;11:2729-2740.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97336071</p> <p>Reference Title: A family of LIM domain-associated cofactors confer transcriptional synergism between LIM and Otx homeodomain proteins.</p> <p>Reference Author: Bach I, Carriere C, Ostendorff HP, Andersen B, Rosenfeld MG;</p> <p>Reference Location: Genes Dev 1997;11:1370-1380</p> <p>Reference Number: [3]</p> <p>Reference Medline: 97078753</p> <p>Reference Title: Interactions of the LIM-domain-binding factor Ldb1 with LIM homeodomain proteins.</p> <p>Reference Author: Aguinick AD, Taira M, Breen JJ, Tanaka T, Dawid IB, Westphal H;</p> <p>Reference Location: Nature 1996;384:270-272</p> <p>Reference Number: [4]</p> <p>Reference Medline: 97030257</p> <p>Reference Title: Nuclear LIM interactor, a rhombotin and LIM homeodomain interacting protein, is expressed early in neuronal development.</p> <p>Reference Author: Jurata LW, Kenny DA, Gill GN,</p> <p>Reference Location: Proc Natl Acad Sci U S A 1996;93:11693-11698.</p> <p>Database Reference: INTERPRO, IPR002691,</p> <p>Comment: The LIM-domain binding protein, binds to the LIM domain LIM of LIM homeodomain proteins which are transcriptional regulators of development.</p> <p>Comment: Nuclear LIM interactor (NLI) / LIM domain-binding protein 1 (LDB1)</p> <p>Comment: Swiss:P70662 is located in the nuclei of neuronal cells during development, it is co-expressed with Isl1 in early motor neuron differentiation and has a suggested role in the Isl1 dependent development of motor neurons [4].</p> <p>Comment: It is suggested that these proteins act synergistically to enhance transcriptional efficiency by acting as co-factors for LIM homeodomain and Otx class transcription factors both of which have essential roles in development [2].</p> <p>Comment: The Drosophila protein Chip Swiss:O18353 is required for segmentation and activity of a remote wing margin enhancer [1]. Chip is a ubiquitous chromosomal factor required for normal expression of diverse genes at many stages of development [1]. It is suggested that Chip cooperates with different LIM domain proteins and other factors to structurally support remote enhancer-promoter interactions [1]</p> <p>Number of members: 19</p>
Lipase_3	PDOC00110	Lipases, serine active site	<p>Triglyceride lipases (EC 3.1.1.3) [1] are lipolytic enzymes that hydrolyzes the ester bond of triglycerides. Lipases are widely distributed in animals, plants and prokaryotes. In higher vertebrates there are at least three tissue-specific isozymes: pancreatic, hepatic, and gastric/lingual. These three types of lipases are closely related to each other as well as to lipoprotein lipase (EC 3.1.1.34) [2], which hydrolyzes triglycerides of chylomicrons and very low</p>

928

			<p>density lipoproteins (VLDL).</p> <p>The most conserved region in all these proteins is centered around a serine residue which has been shown [3] to participate, with an histidine and an aspartic acid residue, to a charge relay system. Such a region is also present in lipases of prokaryotic origin and in lecithin-cholesterol acyltransferase (EC 2.3.1.43) (LCAT) [4], which catalyzes fatty acid transfer between phosphatidylcholine and cholesterol. We have built a pattern from that region.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIV]-x-[LIVFY]-[LIVMST]-G-[HYWV]-S-x-G-[GSTAC] [S is the active site residue] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT 35</p> <p>Note Drosophila vitellogenins are also related to lipases [5], but they have lost their active site serine.</p> <p>Last update November 1997 / Pattern and text revised</p> <p>References [1] Chapus C., Rovey M., Sarda L., Verger R Biochimie 70:1223-1234(1988)</p> <p>[2] Persson B., Bengtsson-Olivecrona G., Enerback S., Olivecrona T., Joernvall H. Eur. J. Biochem. 179:39-45(1989).</p> <p>[3] Blow D. Nature 343:694-695(1990).</p> <p>[4] McLean J., Fielding C., Drayna D., Dieplinger H., Baer B., Kohr W., Henzel W., Lawn R Proc. Natl. Acad. Sci. U.S.A. 83:2335-2339(1986).</p> <p>[5] Baker M E. Biochem. J. 255:1057-1060(1988).</p>
Lipase_GDSL	PDOC00842	Lipolytic enzymes "G-D-S-L" family, serine active site	<p>Recently [1], a family of lipolytic enzymes has been characterized. This family currently consist of the following proteins:</p> <ul style="list-style-type: none"> - Aeromonas hydrophila lipase/phosphatidylcholine-sterol acyltransferase. - Xenorhabdus luminescens lipase 1 - Vibrio mimicus arylesterase. - Escherichia coli acyl-coA thioesterase I (gene tesA). - Vibrio parahaemolyticus thermolabile hemolysin/atypical phospholipase. - Rabbit phospholipase AdRab-B, an intestinal brush border protein with esterase and phospholipase A/lysophospholipase activity that could be involved in the uptake of dietary lipids. AdRab-B contains four repeats of about 320 amino acids. - Arabidopsis thaliana and Brassica napus anther-specific proline-rich protein APG - A Pseudomonas putida hypothetical protein in trpE-trpG intergenic region <p>A serine has been identified a part of the active site in the Aeromonas, Vibrio mimicus and Escherichia coli enzymes. It is located in a conserved sequence motif that can be used as a signature pattern for these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVMFYAG](4)-G-D-S-[LIVM]-x(1,2)-[TAG]-G [S is the active site residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note this pattern will pick up two of the four repeats in AdRab-B. the first one is</p>

			<p>not detected as its sequence has diverged in the region of the putative active site residue. The last one is also not detected because it is slightly divergent at the end of the pattern.</p> <p>Expert(s) to contact by email Upton C. upton@sol.uvic.ca</p> <p>Buckley J.T. tbuckley@sol.uvic.ca</p> <p>Last update November 1995 / First entry.</p> <p>References [1] Upton C., Buckley J.T. <i>Trends Biochem. Sci.</i> 20:178-179(1995).</p>
Lipoprotein_1	PDOC00013	Prokaryotic membrane lipoprotein lipid attachment site	<p>In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):</p> <ul style="list-style-type: none"> - Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp). - <i>Escherichia coli</i> lipoprotein-28 (gene nlpA). - <i>Escherichia coli</i> lipoprotein-34 (gene nlpB). - <i>Escherichia coli</i> lipoprotein nlpC. - <i>Escherichia coli</i> lipoprotein nlpD. - <i>Escherichia coli</i> osmotically inducible lipoprotein B (gene osmB). - <i>Escherichia coli</i> osmotically inducible lipoprotein E (gene osmE). - <i>Escherichia coli</i> peptidoglycan-associated lipoprotein (gene pal). - <i>Escherichia coli</i> rare lipoproteins A and B (genes rplA and rplB). - <i>Escherichia coli</i> copper homeostasis protein cutF (or nlpE). - <i>Escherichia coli</i> plasmids traT proteins. - <i>Escherichia coli</i> Col plasmids lysis proteins. - A number of <i>Bacillus</i> beta-lactamases. - <i>Bacillus subtilis</i> periplasmic oligopeptide-binding protein (gene oppA). - <i>Borrelia burgdorferi</i> outer surface proteins A and B (genes ospA and ospB). - <i>Borrelia hermsii</i> variable major protein 21 (gene vmp21) and 7 (gene vmp7). - <i>Chlamydia trachomatis</i> outer membrane protein 3 (gene omp3). - <i>Fibrobacter succinogenes</i> endoglucanase cel-3. - <i>Haemophilus influenzae</i> proteins Pal and Pcp. - <i>Klebsiella pullulunase</i> (gene pulA). - <i>Klebsiella pullulunase</i> secretion protein pulS. - <i>Mycoplasma hyorhinis</i> protein p37. - <i>Mycoplasma hyorhinis</i> variant surface antigens A, B, and C (genes vlpABC). - <i>Neisseria</i> outer membrane protein H.8. - <i>Pseudomonas aeruginosa</i> lipopeptide (gene lppL). - <i>Pseudomonas solanacearum</i> endoglucanase egl. - <i>Rhodospseudomonas viridis</i> reaction center cytochrome subunit (gene cytC). - <i>Rickettsia</i> 17 Kd antigen. - <i>Shigella flexneri</i> invasion plasmid proteins mxlJ and mxlM. - <i>Streptococcus pneumoniae</i> oligopeptide transport protein A (gene amiA). - <i>Treponema pallidum</i> 34 Kd antigen. - <i>Treponema pallidum</i> membrane protein A (gene tmpA). - <i>Vibrio Harveyi</i> chitobiase (gene chb). - <i>Yersinia</i> virulence plasmid protein yscJ. <p>- Halocyanin from <i>Natrobacterium pharaonis</i> [4]. a membrane associated copper-binding protein. This is the first archaeobacterial protein known to be modified in such a fashion).</p> <p>From the precursor sequences of all these proteins, we derived a consensus pattern and a set of rules to identify this type of post-translational modification.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1) The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least</p>

		<p>one Lys or one Arg in the first seven positions of the sequence. Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT some 100 prokaryotic proteins. Some of them are not membrane lipoproteins, but at least half of them could be. Last update November 1995 / Pattern and text revised. References [1] Hayashi S., Wu H C. J. Bioenerg. Biomembr 22:451-471(1990). [2] Klein P., Somorjai R.L., Lau P.C.K. Protein Eng 2:15-20(1988). [3] von Heijne G. Protein Eng 2:531-534(1989). [4] Mattar S., Scharf B., Kent S B H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem 269:14939-14945(1994).</p>
Lipoprotein_2	PDOC00013	<p>Prokaryotic membrane lipoprotein lipid attachment site</p> <p>In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]):</p> <ul style="list-style-type: none"> - Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp) - Escherichia coli lipoprotein-28 (gene nlpA) - Escherichia coli lipoprotein-34 (gene nlpB). - Escherichia coli lipoprotein nlpC. - Escherichia coli lipoprotein nlpD - Escherichia coli osmotically inducible lipoprotein B (gene osmB). - Escherichia coli osmotically inducible lipoprotein E (gene osmE). - Escherichia coli peptidoglycan-associated lipoprotein (gene pal) - Escherichia coli rare lipoproteins A and B (genes rplA and rplB). - Escherichia coli copper homeostasis protein cutF (or nlpE). - Escherichia coli plasmids traT proteins - Escherichia coli Col plasmids lysis proteins - A number of Bacillus beta-lactamases. - Bacillus subtilis periplasmic oligopeptide-binding protein (gene oppA). - Borrelia burgdorferi outer surface proteins A and B (genes ospA and ospB) - Borrelia hermsii variable major protein 21 (gene vmp21) and 7 (gene vmp7). - Chlamydia trachomatis outer membrane protein 3 (gene omp3). - Fibrobacter succinogenes endoglucanase cel-3 - Haemophilus influenzae proteins Pal and Pcp. - Klebsiella pullulunase (gene pulA). - Klebsiella pullulunase secretion protein pulS. - Mycoplasma hyorhinis protein p37 - Mycoplasma hyorhinis variant surface antigens A, B, and C (genes vlpABC). - Neisseria outer membrane protein H.8. - Pseudomonas aeruginosa lipopeptide (gene lppL). - Pseudomonas solanacearum endoglucanase egl. - Rhodospseudomonas viridis reaction center cytochrome subunit (gene cytC) - Rickettsia 17 Kd antigen. - Shigella flexneri invasion plasmid proteins mxj and mxm. - Streptococcus pneumoniae oligopeptide transport protein A (gene amA). - Treponema pallidum 34 Kd antigen. - Treponema pallidum membrane protein A (gene tmpA). - Vibrio Harveyi chitobiase (gene chb). - Yersinia virulence plasmid protein yscJ. - Halocyanin from Natrobacterium pharaonis [4], a membrane associated copper-binding protein. This is the first archaeobacterial protein known to be modified in such a fashion). <p>From the precursor sequences of all these proteins, we derived a consensus pattern and a set of rules to identify this type of post-translational modification.</p>

			<p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1) The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least one Lys or one Arg in the first seven positions of the sequence. Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT some 100 prokaryotic proteins. Some of them are not membrane lipoproteins, but at least half of them could be. Last update November 1995 / Pattern and text revised. References [1] Hayashi S., Wu H C J. Bioenerg. Biomembr 22:451-471(1990). [2] Klein P., Somorjai R L., Lau P.C.K. Protein Eng. 2:15-20(1988). [3] von Heijne G. Protein Eng 2 531-534(1989). [4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterheld D., Engelhard M. J. Biol. Chem 269:14939-14945(1994).</p>
Lipoprotein_5	PDOC00013	Prokaryotic membrane lipoprotein lipid attachment site	<p>In prokaryotes, membrane lipoproteins are synthesized with a precursor signal peptide, which is cleaved by a specific lipoprotein signal peptidase (signal peptidase II). The peptidase recognizes a conserved sequence and cuts upstream of a cysteine residue to which a glyceride-fatty acid lipid is attached [1]. Some of the proteins known to undergo such processing currently include (for recent listings see [1,2,3]).</p> <ul style="list-style-type: none"> - Major outer membrane lipoprotein (murein-lipoproteins) (gene lpp) - Escherichia coli lipoprotein-28 (gene nlpA). - Escherichia coli lipoprotein-34 (gene nlpB). - Escherichia coli lipoprotein nlpC - Escherichia coli lipoprotein nlpD - Escherichia coli osmotically inducible lipoprotein B (gene osmB). - Escherichia coli osmotically inducible lipoprotein E (gene osmE). - Escherichia coli peptidoglycan-associated lipoprotein (gene pal) - Escherichia coli rare lipoproteins A and B (genes rplA and rplB). - Escherichia coli copper homeostasis protein cutF (or nlpE). - Escherichia coli plasmids traT proteins. - Escherichia coli Col plasmids lysis proteins. - A number of Bacillus beta-lactamases. - Bacillus subtilis periplasmic oligopeptide-binding protein (gene oppA) - Borrelia burgdorferi outer surface proteins A and B (genes ospA and ospB) - Borrelia hermsii variable major protein 21 (gene vmp21) and 7 (gene vmp7). - Chlamydia trachomatis outer membrane protein 3 (gene omp3). - Fibrobacter succinogenes endoglucanase cel-3. - Haemophilus influenzae proteins Pal and Pcp. - Klebsiella pullulunase (gene pulA). - Klebsiella pullulunase secretion protein pulS. - Mycoplasma hyorhinis protein p37. - Mycoplasma hyorhinis variant surface antigens A, B, and C (genes vlpABC) - Neisseria outer membrane protein H.8. - Pseudomonas aeruginosa lipopeptide (gene lppL). - Pseudomonas solanacearum endoglucanase egl. - Rhodospseudomonas viridis reaction center cytochrome subunit (gene cytC) - Rickettsia 17 Kd antigen. - Shigella flexneri invasion plasmid proteins mxiJ and mxiM. - Streptococcus pneumoniae oligopeptide transport protein A (gene amIA) - Treponema pallidum 34 Kd antigen. - Treponema pallidum membrane protein A (gene tmpA) - Vibrio harveyi chitobiase (gene chb). - Yersinia virulence plasmid protein yscJ.

932

			<p>- Halocyanin from <i>Natrobacterium pharaonis</i> [4], a membrane associated copper-binding protein. This is the first archaeobacterial protein known to be modified in such a fashion).</p> <p>From the precursor sequences of all these proteins, we derived a consensus pattern and a set of rules to identify this type of post-translational modification.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern {DERK}(6)-[LIVMFWSTAG](2)-[LIVMFYSTAGCQ]-[AGS]-C [C is the lipid attachment site] Additional rules: 1) The cysteine must be between positions 15 and 35 of the sequence in consideration. 2) There must be at least one Lys or one Arg in the first seven positions of the sequence. Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT some 100 prokaryotic proteins. Some of them are not membrane lipoproteins, but at least half of them could be.</p> <p>Last update November 1995 / Pattern and text revised.</p> <p>References [1] Hayashi S., Wu H.C J. Bioenerg. Biomembr. 22:451-471(1990).</p> <p>[2] Klein P., Somorjai R.L., Lau P.C K Protein Eng. 2:15-20(1988).</p> <p>[3] von Heijne G. Protein Eng. 2:531-534(1989).</p> <p>[4] Mattar S., Scharf B., Kent S.B.H., Rodewald K., Oesterhelt D., Engelhard M. J. Biol. Chem. 269:14939-14945(1994).</p>
Luteo_Vpg		Luteovirus putative VPg genome linked protein	<p>Accession number: PF01659</p> <p>Definition: Luteovirus putative VPg genome linked protein</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_970 (release 4.1)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 191.70 191 70</p> <p>Noise cutoffs: -47 90 -47 90</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 94120742</p> <p>Reference Title: Soybean dwarf luteovirus contains the third variant genome</p> <p>Reference Title: type in the luteovirus group.</p> <p>Reference Author: Rathjen JP, Karageorgos LE, Habili N, Waterhouse PM, Symons</p> <p>Reference Author: RH,</p> <p>Reference Location: Virology 1994,198:671-679.</p> <p>Database Reference: INTERPRO, IPR001964,</p> <p>Comment: This family consists of several putative genome linked proteins.</p> <p>Comment: The genomic RNA of luteoviruses are linked to virally encoded genome</p> <p>Comment: proteins (VPg). Open reading frame 4 is thought to encode the VPg</p> <p>Comment: in Soybean dwarf luteovirus [1].</p> <p>Comment: Luteoviruses have isometric capsids that contain a positive strand</p> <p>Comment: ssRNA genome, they have no DNA stage during their replication.</p> <p>Number of members: 32</p>
MATH		MATH domain	<p>Accession number: PF00917</p> <p>Definition: MATH domain</p>

933

			<p>Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_1602 (release 3.0) Gathering cutoffs: 17 0 Trusted cutoffs: 17 90 0.20 Noise cutoffs: 11 80 11.80 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmlcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 96334294 Reference Title: TRAF proteins and meprins share a conserved domain. Reference Author: Uren AG, Vaux DL, Reference Location: Trends Biochem Sci 1996;21 244-245. Reference Number: [2] Reference Medline: 99342031 Reference Title: Crystallographic analysis of CD40 recognition and signaling Reference Title: by human TRAF2. Reference Author: McWhirter SM, Pullen SS, Holton JM, Crute JJ, Kehry MR, Reference Author: Alber T; Reference Location: Proc Natl Acad Sci U S A 1999;96:8408-8413. Reference Number: [3] Reference Medline: 99069615 Reference Title: Comparison of the complete protein sets of worm and yeast: Reference Title: orthology and divergence Reference Author: Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Reference Author: Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T, Weng S, Reference Author: Cherry JM, Botstein D; Reference Location: Science 1998;282 2022-2028. Database Reference: SCOP; 1qsc; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR002083; Database Reference: PDB; 1qsc A; 357; 498; Database Reference: PDB; 1qsc B; 357; 498; Database Reference: PDB; 1qsc C; 357; 498; Database reference: PFAMB; PB018448; Database reference: PFAMB; PB040690; Database reference: PFAMB; PB041198; Comment: This motif has been called the Meprin And TRAF-Homology Comment: (MATH) domain. This domain is hugely expanded in the nematode Comment: C. elegans [3] Number of members: 212</p>
MCT		Monocarboxylate transporter	<p>Accession number: PF01587 Definition: Monocarboxylate transporter Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_483 (release 4.1) Gathering cutoffs: 25 25 Trusted cutoffs: 322.90 322 90 Noise cutoffs: -38 20 -38 20 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmlcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98087501 Reference Title: Cloning and sequencing of four new mammalian monocarboxylate transporter (MCT) homologues confirms the Reference Title: existence of a transporter family with an ancient past. Reference Author: Price NT, Jackson VN, Halestrap AP, Reference Location: Biochem J 1998;329:321-328. Database Reference: INTERPRO; IPR002897; Comment: This domain consists of the transmembrane region of the monocarboxylate Comment: transporters Monocarboxylate transporters (MTC) are transmembrane Comment: glycoproteins with 10-12 predicted transmembrane regions. Comment: They catalyse the proton linked transport of lactic acid,</p>

			<p>Comment: pyruvate and ketone bodies across the plasma membrane [1]</p> <p>Number of members: 33</p>
Methionine_synth		Methionine synthase, vitamin-B12 independent	<p>Accession number: PF01717</p> <p>Definition: Methionine synthase, vitamin-B12 independent</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1909 (release 4.1)</p> <p>Gathering cutoffs: -155.0 -155.0</p> <p>Trusted cutoffs: -155.00 -155.00</p> <p>Noise cutoffs: -170.00 -170.00</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98301657</p> <p>Reference Title: The specific features of methionine biosynthesis and metabolism in plants.</p> <p>Reference Author: Ravanel S, Gakiere B, Job D, Douce R;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1998;95:7805-7812.</p> <p>Database Reference: INTERPRO; IPR002629,</p> <p>Database reference: PFAMB; PB041617;</p> <p>Comment: This is a family of vitamin-B12 independent methionine synthases</p> <p>Comment: or 5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferases, EC.2.1.1 14 from bacteria and plants.</p> <p>Comment: Plants are the only higher eukaryotes that have the required enzymes</p> <p>Comment: for methionine synthesis [1]</p> <p>Comment: This enzyme catalyses the last step in the production of methionine</p> <p>Comment: by transferring a methyl group from 5-methyltetrahydrofolate to</p> <p>Comment: homocysteine [1].</p> <p>Comment: The aligned region makes up the carboxy region of the approximately</p> <p>Comment: 750 amino acid protein except in some hypothetical archaeal proteins</p> <p>Comment: present in the family, where this region corresponds to the entire length.</p> <p>Comment:</p> <p>Number of members: 28</p>
Methyltransf_2		O-methyltransferase	<p>Accession number: PF00891</p> <p>Definition: O-methyltransferase</p> <p>Previous Pfam IDs: Methyltransf;</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_152 (release 3.0)</p> <p>Gathering cutoffs: -53 -53</p> <p>Trusted cutoffs: -22.00 -22.00</p> <p>Noise cutoffs: -84.60 -84.60</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 93167811</p> <p>Reference Title: Purification of a 40-kilodalton methyltransferase active in the aflatoxin biosynthetic pathway.</p> <p>Reference Title:</p> <p>Reference Author: Keller NP, Dischinger HC, Bhatnagar D, Cleveland TE, Ullah</p> <p>Reference Author: AH;</p> <p>Reference Location: Appl Environ Microbiol 1993;59:479-484.</p> <p>Database Reference: INTERPRO: IPR001077;</p> <p>Comment: This family includes a range of O-methyltransferases. These</p> <p>Comment: enzymes utilise S-adenosyl methionine.</p> <p>Number of members: 67</p>
Methyltransf_3		O-methyltransferase	<p>Accession number: PF01596</p> <p>Definition: O-methyltransferase</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_749 (release 4.1)</p> <p>Gathering cutoffs: -86 -86</p> <p>Trusted cutoffs: -81.80 -81.80</p>

935

			<p>Noise cutoffs: -91.00 -91.00 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 97090395 Reference Title: Two multifunctional peptide synthetases and an Reference Title: O-methyltransferase are involved in the biosynthesis of the Reference Title: DNA-binding antibiotic and antitumour agent saframycin Mx1 Reference Title: from Myxococcus xanthus Reference Author: Pospiech A, Bietenhader J, Schupp T; Reference Location: Microbiology 1996;142:741-746 Database Reference: SCOP; 1vid; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR002935; Database Reference: PDB; 1vid; 13; 186; Database reference: PFAMB; PB040269; Comment: Members of this family are O-methyltransferases. The family Comment: includes catechol o-methyltransferase Swiss:P21964, caffeoyl-CoA Comment: O-methyltransferase Swiss:Q43095 and a family of bacterial Comment: O-methyltransferases that may be involved in antibiotic Comment: production [1] Number of members: 39</p>
MMR_HSR1		GTPase of unknown function	<p>Accession number: PF01926 Definition: GTPase of unknown function Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: -21 -21 Trusted cutoffs: -20.70 -20.70 Noise cutoffs: -31.60 -31.60 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 94235953 Reference Title: Structure and evolution of a member of a new subfamily of Reference Title: GTP-binding proteins mapping to the human MHC class I region. Reference Title: Vernet C, Ribouchon MT, Chimini GPontarotti P; Reference Location: Mamm Genome 1994;5:100-105. Database Reference: INTERPRO; IPR002917; Database reference: PFAMB; PB000471; Database reference: PFAMB; PB002171; Database reference: PFAMB; PB015790; Number of members: 67</p>
MoaC		MoaC family	<p>Accession number: PF01967 Definition: MoaC family Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: 25 25 Trusted cutoffs: 73 00 73 00 Noise cutoffs: -93 90 -93.90 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmbuild --seed 0 HMM Reference Number: [1] Reference Medline: 99337076 Reference Title: Characterization of a molybdenum cofactor biosynthetic gene Reference Title: cluster in Rhodobacter capsulatus which is specific for the Reference Title: biogenesis of dimethylsulfoxide reductase Reference Author: Solomon PS, Shaw AL, Lane I, Hanson GR, Palmer T, McEwan Reference Author: AG; Reference Location: Microbiology 1999;145:1421-1429. Database Reference: INTERPRO; IPR002820; Comment: Members of this family are involved in molybdenum Comment: cofactor biosynthesis. However their molecular</p>

936

			<p>Comment: function is not known.</p> <p>Number of members: 24</p>
Myc_N_term		Myc amino-terminal region	<p>Accession number: PF01056</p> <p>Definition: Myc amino-terminal region</p> <p>Author: Finn RD, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_387 (release 3 0)</p> <p>Gathering cutoffs: -109 -109</p> <p>Trusted cutoffs: -81.20 -81.20</p> <p>Noise cutoffs: -137.40 -137.40</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98280742</p> <p>Reference Title: The molecular role of Myc in growth and transformation</p> <p>Reference Title: recent discoveries lead to new insights.</p> <p>Reference Author: Facchini LM, Penn LZ,</p> <p>Reference Location: FASEB J 1998;12 633-651.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97318600</p> <p>Reference Title: Myc target genes.</p> <p>Reference Author: Grandori C, Eisenman RN,</p> <p>Reference Location: Trends Biochem Sci 1997;22 177-181.</p> <p>Database Reference: INTERPRO; IPR002418,</p> <p>Comment: The myc family belongs to the basic helix-loop-helix leucine zipper</p> <p>Comment: class of transcription factors. see HLH. Myc forms a</p> <p>Comment: heterodimer with Max, and this complex regulates cell growth through</p> <p>Comment: direct activation of genes involved in cell replication [2].</p> <p>Number of members: 56</p>
Myosin_tail		Myosin tail	<p>Accession number: PF01576</p> <p>Definition: Myosin tail</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_356 (release 4 1)</p> <p>Gathering cutoffs: 19 19</p> <p>Trusted cutoffs: 23.30 23.30</p> <p>Noise cutoffs: 15.10 15.10</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 87060988</p> <p>Reference Title: Complete nucleotide and encoded amino acid sequence of a</p> <p>Reference Title: mammalian myosin heavy chain gene. Evidence against</p> <p>Reference Title: intron-dependent evolution of the rod.</p> <p>Reference Author: Strehler EE, Strehler-page M-A, Perriard JC, Periasamy M,</p> <p>Reference Author: Nadal-ginard B;</p> <p>Reference Location: J MOL BIOL 1986;190:291-317.</p> <p>Database Reference: INTERPRO, IPR002928,</p> <p>Comment: The myosin molecule is a multi-subunit complex made up of two heavy chains and four light chains it is a</p> <p>Comment: fundamental contractile</p> <p>Comment: protein found in all eukaryote cell types [1].</p> <p>Comment: This family consists of the coiled-coil myosin heavy chain tail region.</p> <p>Comment: The coiled-coil is composed of the tail from two molecules of myosin.</p> <p>Comment: These can then assemble into the macromolecular thick filament [1]</p> <p>Comment: The coiled-coil region provides the structural backbone the thick</p> <p>Comment: filament [1].</p> <p>Number of members: 182</p>
Na_Ala_symp	PDOC00681	Sodium:alanine symporter family signature	<p>It has been shown [1] that integral membrane proteins that mediate the intake of a wide variety of molecules with the concomitant uptake of sodium ions (sodium symporters) can be grouped, on the basis of sequence and functional similarities into a number of distinct families. One of these families is known as the sodium:alanine symporter family (SAF) and currently consists of</p>

937

			<p>the following proteins:</p> <ul style="list-style-type: none"> - Thermophilic bacterium PS-3 alanine carrier protein (ACP). ACP can use both sodium and hydrogen as a symport ion - Alteromonas haloplanktis D-alanine/glycine permease (gene dagA). - Bacillus subtilis alsT. - Hypothetical protein yaaJ from Escherichia coli and HI0183, the corresponding Haemophilus influenzae protein. - Haemophilus influenzae hypothetical protein HI0883. <p>These integral membrane proteins are predicted to comprise a least eight membrane spanning domains. As a signature pattern we selected a highly conserved region which is located in the N-terminal section and which includes part of the first transmembrane region.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-G-x-[GA](2)-[LIVM]-F-W-M-W-[LIVM]-x-[STAV]-[LIVMFA](2)-G</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Pattern and text revised</p> <p>References [1] Reizer J , Reizer A., Saier M H. Jr. Biochim. Biophys. Acta 1197:133-136(1994).</p>
Na_Ca_Ex		Sodium/calcium exchanger protein	<p>Accession number: PF01699</p> <p>Definition: Sodium/calcium exchanger protein</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1680 (release 4.1)</p> <p>Gathering cutoffs: 3 3</p> <p>Trusted cutoffs: 3.40 3 40</p> <p>Noise cutoffs: 1.20 1.20</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96394663</p> <p>Reference Title: Cloning of a third mammalian Na⁺-Ca²⁺ exchanger, NCX3</p> <p>Reference Author: Nicoll DA, Quednau BD, Qui Z, Xia YR, Lysis AJ, Philipson</p> <p>Reference Author KD:</p> <p>Reference Location: J Biol Chem 1996;271:24914-24921.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 91047958</p> <p>Reference Title: Molecular cloning and functional expression of the cardiac sarcolemmal Na⁺(+)-Ca²⁺ exchanger.</p> <p>Reference Author: Nicoll DA, Longoni S, Philipson KD;</p> <p>Reference Location: Science 1990;250:562-565.</p> <p>Database Reference: INTERPRO, IPR002613,</p> <p>Database reference: PFAMB; PB002768;</p> <p>Database reference: PFAMB; PB040773;</p> <p>Database reference: PFAMB; PB041540;</p> <p>Comment: This is a family of sodium/calcium exchanger integral membrane proteins. This family covers the integral membrane regions of the proteins. Sodium/calcium exchangers regulate intracellular Ca²⁺ concentrations in many cells; cardiac myocytes, epithelial cells, neurons retinal rod photoreceptors and smooth muscle cells [2].</p> <p>Comment: Ca²⁺ is moved into or out of the cytosol depending on Na⁺ concentration [2]. In humans and rats there are 3 isoforms; NCX1 NCX2 and NCX3 [1]</p> <p>Comment: see Swiss.Q01728, Swiss:P48768 and Swiss:P70549</p>

			respectively. Number of members: 105
Na_K_ATPase_C		Na+/K+ ATPase C-terminus	<p>This domain is specific to the sodium and potassium ATPases (Na_K-ATPase). The sodium pump (Na+,K+ ATPase), located in the plasma membrane of all animal cells [1], is an heterotrimer of a catalytic subunit (alpha chain), a glycoprotein subunit of about 34 Kd (beta chain) and a small hydrophobic protein of about 6 Kd. The beta subunit seems [2] to regulate, through the assembly of alpha/beta heterodimers, the number of sodium pumps transported to the plasma membrane.</p> <p>This family is typically found in association with E1-E2 ATPase. Uses of these polypeptide includes regulating that ion content in a desired cell or organism and can convey salt or ion tolerance.</p>
Na_K_ATPase_N		Na+/K+ ATPase C-terminus	<p>Accession number: PF00690 Definition: Na+/K+ ATPase C-terminus Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_138 (release 2 1) Gathering cutoffs: 15 6 15 6 Trusted cutoffs: 15.60 15 60 Noise cutoffs: 15.10 15.10 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Database Reference INTERPRO: IPR000661; Database reference: PFAMB: PB000031; Comment: This family is always found in association with E1-E2_ATPase. Comment: This extension is specific to the Na+/K+ ATPase subfamily of Comment: ATPases. Number of members: 90</p>
NAD_Gly3P_dh	PDOC00740	NAD-dependent glycerol-3-phosphate dehydrogenase signature	<p>NAD-dependent glycerol-3-phosphate dehydrogenase (EC 1.1.1.8) (GPD) catalyzes the reversible reduction of dihydroxyacetone phosphate to glycerol-3-phosphate. It is a eukaryotic cytosolic homodimeric protein of about 40 Kd. As a signature pattern we selected a glycine-rich region that is probably [1] involved in NAD-binding.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-[AT]-[LIVM]-K-[DN]-[LIVM](2)-A-x-[GA]-x-G-[LIVMF]-x-[DE]-G-[LIVM]-x-[LIVMFYW]-G-x-N Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update November 1997 / Pattern and text revised. References [1] Otto J., Argos P , Rossmann M.G. Eur. J. Biochem 109 325-330(1980).</p>
NifU_N		NifU-like N terminal domain	<p>Accession number: PF01592 Definition: NifU-like N terminal domain Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_772 (release 4 1) Gathering cutoffs: -13 -13 Trusted cutoffs: 1.20 1.20 Noise cutoffs: -28 80 -28.80 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97032601 Reference Title: A modular domain of NifU, a nitrogen fixation cluster protein, is highly conserved in evolution. Reference Author: Hwang DM, Dempsey A, Tan KT, Liew CC; Reference Location: J Mol Evol 1996;43:536-540. Database Reference INTERPRO; IPR002871, Comment: This domain is found in NifU in combination with NifU-like. Comment: This domain is found on isolated in several bacterial</p>

			<p>species</p> <p>Comment: such as Swiss:O53156. The nif genes are responsible for nitrogen</p> <p>Comment: fixation. However this domain is found in bacteria that do not</p> <p>Comment: fix nitrogen, so it may have a broader significance in the cell</p> <p>Comment: than nitrogen fixation.</p> <p>Number of members: 32</p>
NTR		NTR/C345C module	<p>Accession number PF01759</p> <p>Definition: NTR/C345C module</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: [1]</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 57.30 57.30</p> <p>Noise cutoffs: 2.80 2.80</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 99379676</p> <p>Reference Title: The NTR module: domains of netrins, secreted frizzled related proteins, and type I procollagen C-proteinase enhancer protein are homologous with tissue inhibitors of metalloproteases [in Process Citation]</p> <p>Reference Author: Banyai L, Pathy L;</p> <p>Reference Location: Protein Sci 1999;8:1636-1642.</p> <p>Database Reference: INTERPRO, IPR001134;</p> <p>Database reference: PFAM; PB005955,</p> <p>Comment: We have not included the related TIMP family.</p> <p>Comment: It has been suggested that the common function of these modules is binding to metzincins [1]. A subset of this family</p> <p>Comment: is known as the C345C domain because it occurs in complement</p> <p>Comment: C3, C4 and C5.</p> <p>Number of members: 64</p>
Nucleoside_transporter		Nucleoside transporter	<p>Accession number: PF01733</p> <p>Definition: Nucleoside transporter</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_2135 (release 4.1)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 25.50 25.50</p> <p>Noise cutoffs: -122.50 -122.50</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98148080</p> <p>Reference Title: Cloning of the human equilibrative, nitrobenzylmercaptapurine riboside (NBMPR)-insensitive nucleoside transporter ei by functional expression in a transport-deficient cell line.</p> <p>Reference Author: Crawford CR, Patel DH, Naeve C, Belt JA;</p> <p>Reference Location: J Biol Chem 1998;273:5288-5293.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 98019212</p> <p>Reference Title: Molecular cloning and functional characterization of nitrobenzylthionosine (NBMPR)-sensitive (es) and NBMPR-insensitive (ei) equilibrative nucleoside transporter</p> <p>Reference Title: proteins (rENT1 and rENT2) from rat tissues</p> <p>Reference Author: Yao SY, Ng AM, Muzyka WR, Griffiths M, Cass CE, Baldwin SA,</p> <p>Reference Author: Young JD;</p> <p>Reference Location: J Biol Chem 1997;272:28423-28430.</p> <p>Database Reference: INTERPRO; IPR002259;</p> <p>Comment: This is a family of nucleoside transporters.</p> <p>Comment: In mammalian cells nucleoside transporters transport nucleoside</p> <p>Comment: across the plasma membrane and are essential for nucleotide</p>

940

			<p>Comment: synthesis via the salvage pathways for cells that lack their own</p> <p>Comment: de novo synthesis pathways [2].</p> <p>Comment: Also in this family is mouse and human nucleolar protein HNP36</p> <p>Comment: Swiss:Q14542 a protein of unknown function, although it has been</p> <p>Comment: hypothesized to be a plasma membrane nucleoside transporter [2].</p> <p>Number of members: 15</p>
Orbi_VP6		Orbivirus helicase VP6	<p>Accession number: PF01516</p> <p>Definition: Orbivirus helicase VP6</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_765 (release 4.0)</p> <p>Gathering cutoffs: -68 -68</p> <p>Trusted cutoffs: -37.10 -37.10</p> <p>Noise cutoffs: -98.90 -98.90</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97456481</p> <p>Reference Title: Bluetongue virus VP6 protein binds ATP and exhibits an RNA-dependent ATPase function and a helicase activity that</p> <p>Reference Title: catalyze the unwinding of double-stranded RNA substrates.</p> <p>Reference Author: Stauber N, Martinez-Costas J, Sutton G, Monastyrskaya K,</p> <p>Reference Author: Roy P;</p> <p>Reference Location: J Virol 1997;71:7220-7226.</p> <p>Database Reference: INTERPRO; IPR001399;</p> <p>Comment: The VP6 protein a minor protein in the core of the virion</p> <p>Comment: is probably the viral helicase [1].</p> <p>Number of members: 27</p>
OSCP	PDOC00327	ATP synthase delta (OSCP) subunit signature	<p>ATP synthase (proton-translocating ATPase) (EC 3.6.1.34) [1,2] is a component of the cytoplasmic membrane of eubacteria, the inner membrane of mitochondria, and the thylakoid membrane of chloroplasts. The ATPase complex is composed of an oligomeric transmembrane sector, called CF(0), which acts as a proton channel, and a catalytic core, termed coupling factor CF(1)</p> <p>One of the subunits of the ATPase complex, known as subunit delta in bacteria and chloroplasts or the Oligomycin Sensitivity Conferral Protein (OSCP) in mitochondria, seems to be part of the stalk that links CF(0) to CF(1). It either transmits conformational changes from CF(0) into CF(1) or is involved in proton conduction [3].</p> <p>The different delta/OSCP subunits are proteins of approximately 200 amino-acid residues - once the transit peptide has been removed in the chloroplast and mitochondrial forms - which show only moderate sequence homology.</p> <p>The signature pattern used to detect ATPase delta/OSCP subunits is based on a conserved region in the C-terminal section of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVM]-x-[LIVMFYT]-x(3)-[LIVMT]-[DENQK]-x(2)-[LIVM]-x-[GSA]-G-[LIVMFYGA]-x-[LIVM]-[KRHENQ]-x-[GSEN]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except 3 sequences</p> <p>Other sequence(s) detected in SWISS-PROT 2.</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References [1]</p> <p>Futai M., Noumi T., Maeda M.</p>

			<p>Annu. Rev. Biochem. 58:111-136(1989).</p> <p>[2] Senior A.E. Physiol. Rev. 68:177-231(1988).</p> <p>[3] Engelbrecht S., Junge W. Biochim. Biophys. Acta 1015:379-390(1990).</p>
OTCace	PDOC00091	Aspartate and ornithine carbamoyltransferases signature	<p>Aspartate carbamoyltransferase (EC 2.1.3.2) (ATCase) catalyzes the conversion of aspartate and carbamoyl phosphate to carbamoylaspartate. the second step in the de novo biosynthesis of pyrimidine nucleotides [1] In prokaryotes ATCase consists of two subunits: a catalytic chain (gene pyrB) and a regulatory chain (gene pyrI), while in eukaryotes it is a domain in a multi-functional enzyme (called URA2 in yeast, rudimentary in Drosophila, and CAD in mammals [2]) that also catalyzes other steps of the biosynthesis of pyrimidines.</p> <p>Ornithine carbamoyltransferase (EC 2.1.3.3) (OTCase) catalyzes the conversion of ornithine and carbamoyl phosphate to citrulline. In mammals this enzyme participates in the urea cycle [3] and is located in the mitochondrial matrix. In prokaryotes and eukaryotic microorganisms it is involved in the biosynthesis of arginine. In some bacterial species it is also involved in the degradation of arginine [4] (the arginine deaminase pathway)</p> <p>It has been shown [5] that these two enzymes are evolutionary related. The predicted secondary structure of both enzymes are similar and there are some regions of sequence similarities. One of these regions includes three residues which have been shown, by crystallographic studies [6], to be implicated in binding the phosphoryl group of carbamoyl phosphate. We have selected this region as a signature for these enzymes.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern F-x-[EK]-x-S-[GT]-R-T [S, R, and the 2nd T bind carbamoyl phosphate] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note the residue in position 3 of the pattern allows to distinguish between an ATCase (Glu) and an OTCase (Lys) Last update October 1993 / Text revised. References [1] Lerner C.G., Switzer R L. J. Biol. Chem. 261:11156-11165(1986).</p> <p>[2] Davidson J.N., Chen K C., Jamison R S., Musmanno L.A , Kern C.B. BioEssays 15:157-164(1993).</p> <p>[3] Takiguchi M., Matsubasa T , Amaya Y , Mori M. BioEssays 10:163-166(1989).</p> <p>[4] Baur H., Stalon V., Falmagne P., Luethi E , Haas D. Eur. J Biochem. 166:111-117(1987).</p> <p>[5] Houghton J.E., Bencini D.A O'Donovan G.A., Wild J.R. Proc. Natl. Acad. Sci. U.S.A. 81:4864-4868(1981).</p> <p>[6] Ke H.-M., Honzatko R.B., Lipscomb W.N. Proc. Natl Acad. Sci. U.S.A. 81:4037-4040(1984).</p>

942

oxidored_q1_N		NADH-Ubiquinone oxidoreductase (complex I), chain 5 N-terminus	<p>Accession number PF00662</p> <p>Definition: NADH-Ubiquinone oxidoreductase (complex I), chain 5 N-terminus</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_22 (release 2.1)</p> <p>Gathering cutoffs: 18 18</p> <p>Trusted cutoffs: 19.40 19.40</p> <p>Noise cutoffs: 16.70 16.70</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 93110040</p> <p>Reference Title: The NADH:ubiquinone oxidoreductase (complex I) of respiratory chains</p> <p>Reference Author: Walker JE;</p> <p>Reference Location: Q Rev Biophys 1992;25 253-324.</p> <p>Database Reference: INTERPRO: IPR001516,</p> <p>Database reference: PFAMB; PB000410;</p> <p>Database reference: PFAMB; PB033295;</p> <p>Database reference: PFAMB; PB040550;</p> <p>Comment: This sub-family represents an amino terminal extension of oxidored_q1 Only NADH-Ubiquinone chain 5 and eubacterial chain L are in this family</p> <p>Comment: This sub-family is part of complex I which catalyses the transfer of two electrons from NADH to ubiquinone in a reaction that is associated with proton translocation across the membrane.</p> <p>Number of members: 546</p>
oxidored_q2		NADH-ubiquinone/plastoquinone oxidoreductase chain 4L	<p>Accession number PF00420</p> <p>Definition: NADH-ubiquinone/plastoquinone oxidoreductase chain 4L</p> <p>Author: Finn RD</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_193 (release 1.0)</p> <p>Gathering cutoffs: 25 15</p> <p>Trusted cutoffs: 29.70 29.70</p> <p>Noise cutoffs: 20.40 20.40</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Database Reference: INTERPRO, IPR001133;</p> <p>Database reference: PFAMB; PB006066;</p> <p>Number of members: 219</p>
PAN	PDOC00376	Apple domain	<p>Plasma kallikrein (EC 3.4.21.34) and coagulation factor XI (EC 3.4.21.27) are two related plasma serine proteases activated by factor XIa and which share the same domain topology: an N-terminal region that contains four tandem repeats of about 90 amino acids and a C-terminal catalytic domain.</p> <p>The 90 amino-acid repeated domain contains 6 conserved cysteines. It has been shown [1,2] that three disulfide bonds link the first and sixth, second and fifth, and third and fourth cysteines. The domain can be drawn in the shape of an apple (see below) and has been accordingly called the 'apple domain'.</p> <pre> x x x x x x x C---C x x x x x x Cx x x x x x x x x Cx x x x x x x x x x x x x x x x x x x x x x x C---C x x... </pre> <p>Schematic representation of an apple domain.</p> <p>Apart from the cysteines, there are a number of other conserved positions in the apple domain. We have developed a pattern, that spans the complete domain, and which includes these conserved positions</p>

			<p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-x(3)-[LIVMFY]-x(5)-[LIVMFY]-x(3)-[DENQ]-[LIVMFY]-x(10)- C-x(3)-C-T-x(4)-C-x-[LIVMFY]-F-x-[FY]-x(13,14)-C-x- [LIVMFY]-[RK]-x-[ST]-x(14,15)-S-G-x-[ST]-[LIVMFY]-x(2)-C</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE</p> <p>Last update June 1992 / Pattern and text revised.</p> <p>References [1] McMullen B.A., Fujikawa K., Davie E.W. Biochemistry 30 2050-2056(1991).</p> <p>[2] McMullen B.A., Fujikawa K., Davie E.W. Biochemistry 30.2056-2060(1991).</p>
PAP2		PAP2 superfamily	<p>Accession number: PF01569</p> <p>Definition: PAP2 superfamily</p> <p>Author: Bashton M. Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_486 (release 4.0)</p> <p>Gathering cutoffs: 16 16</p> <p>Trusted cutoffs: 22.00 22.00</p> <p>Noise cutoffs: 11.40 11.40</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97194074</p> <p>Reference Title: Identification of a novel phosphatase sequence motif</p> <p>Reference Author: Stukey J, Carman GM;</p> <p>Reference Location: Protein Sci 1997;6:469-472.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97406916</p> <p>Reference Title: An unexpected structural relationship between integral membrane phosphatases and soluble haloperoxidases.</p> <p>Reference Author: Neuwald AF;</p> <p>Reference Location: Protein Sci 1997;6:1764-1767.</p> <p>Database Reference: INTERPRO; IPR000326;</p> <p>Database reference: PFAMB; PB021113;</p> <p>Database reference: PFAMB; PB040926;</p> <p>Database reference: PFAMB; PB041096;</p> <p>Database reference: PFAMB; PB041301;</p> <p>Comment: This family includes the enzyme type 2 phosphatidic acid</p> <p>Comment: phosphatase (PAP2).</p> <p>Number of members: 49</p>
PAPS_reduct		Phosphoadenosine phosphosulfate reductase family	<p>Accession number: PF01507</p> <p>Definition: Phosphoadenosine phosphosulfate reductase family</p> <p>Author: Bashton M. Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_590 (release 4.0)</p> <p>Gathering cutoffs: 49 49</p> <p>Trusted cutoffs: 55.40 55.40</p> <p>Noise cutoffs: -34.60 -34.60</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97411695</p> <p>Reference Title: Crystal structure of phosphoadenylyl sulphate (PAPS) reductase: a new family of adenine nucleotide alpha hydrolases.</p> <p>Reference Title: hydrolases.</p> <p>Reference Author: Savage H, Montoya G, Svensson C, Schwenn JD, Sinning I;</p> <p>Reference Location: Structure 1997;5:895-906.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 96061968</p> <p>Reference Title: Reaction mechanism of thioredoxin:</p> <p>Reference Title: 3'-phospho-adenylylsulfate reductase investigated by site-directed mutagenesis.</p> <p>Reference Title: site-directed mutagenesis.</p> <p>Reference Author: Berendt U, Haverkamp T, Prior A. Schwenn JD;</p>

944

			<p>Reference Location: Eur J Biochem 1995;233:347-356. Reference Number: [3] Reference Medline: 91066949 Reference Title: ATP sulphurylase activity of the nodP and nodQ gene products of Rhizobium meliloti. Reference Author: Schwedock J, Long SR; Reference Location: Nature 1990;348:644-647. Database Reference: SCOP, 1sur; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO: IPR002500; Database Reference: PDB, 1sur ; 48; 215; Comment: This domain is found in phosphoadenosine phosphosulfate (PAPS) reductase Comment: enzymes or PAPS sulfotransferase PAPS reductase is part of the adenine Comment: nucleotide alpha hydrolases superfamily also including N type ATP PPases Comment: and ATP sulphurylases [1]. The enzyme uses thioredoxin as an electron Comment: donor for the reduction of PAPS to phospho-adenosine-phosphate (PAP) [1,2]. Comment: It is also found in NodP nodulation protein P from Rhizobium which has ATP Comment: sulphurylase activity (sulfate adenylate transferase) [3]. Number of members: 48</p>
PARP		Poly(ADP-ribose) polymerase catalytic region	<p>Accession number: PF00644 Definition: Poly(ADP-ribose) polymerase catalytic region. Author: Bateman A Alignment method of seed: HMM_built_from_alignment Source of seed members: Bateman A Gathering cutoffs: -59.4 -59.4 Trusted cutoffs: -44.60 -44.60 Noise cutoffs: -180.60 -180.60 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 96353841 Reference Title: Structure of the catalytic fragment of poly(AD-ribose) polymerase from chicken. Reference Title: polymerase from chicken. Reference Author: Ruf A, Mennissier de Murcia J, de Murcia G, Schulz GE; Reference Location: Proc Natl Acad Sci U S A 1996;93:7481-7485. Reference Number: [2] Reference Medline: 93293867 Reference Title: The carboxyl-terminal domain of human poly(ADP-ribose) polymerase. Overproduction in Escherichia coli, large scale Reference Title: purification, and characterization. Reference Author: Simonin F, Hofferer L, Panzeter PL, Muller S, de Murcia G. Reference Author: Althaus FR, Reference Location: J Biol Chem 1993;268:13454-13461. Database Reference: SCOP; 1paw; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR001290, Database Reference: PDB; 1a26 ; 662; 997; Database Reference: PDB; 1pax ; 662; 997; Database Reference: PDB; 2pax ; 662; 997; Database Reference: PDB; 3pax ; 662; 997; Database Reference: PDB; 4pax ; 662; 997; Database Reference: PDB, 2paw ; 662; 1009; Database reference: PFAMB; PB041409; Comment: Poly(ADP-ribose) polymerase catalyses the covalent attachment of ADP-ribose units from NAD+ to itself and to a limited number of other DNA binding proteins, which decreases their affinity for DNA. Comment: Poly(ADP-ribose) polymerase is a regulatory component induced by DNA damage Comment: The carboxyl-terminal region is the most highly conserved region of the protein. Experiments have shown that a carboxyl 40 kDa fragment is still catalytically active [2]. Number of members: 19</p>
PC_rep		Proteasome/cyclosome	<p>Accession number: PF01851 Definition: Proteasome/cyclosome repeat</p>

945

		repeat	<p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: [1]</p> <p>Gathering cutoffs: 25 0</p> <p>Trusted cutoffs: 30 60 3.00</p> <p>Noise cutoffs: 15.80 15.80</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97348748</p> <p>Reference Title: A repetitive sequence in subunits of the 26S proteasome and</p> <p>Reference Title: 20S cyclosome (anaphase-promoting complex).</p> <p>Reference Author: Lupas A, Baumeister W, Hofmann K;</p> <p>Reference Location: Trends Biochem Sci 1997;22:195-196.</p> <p>Database Reference: INTERPRO, IPR002015,</p> <p>Database reference: PFAMB; PB009978;</p> <p>Database reference: PFAMB; PB040656;</p> <p>Number of members: 112</p>
PE		PE family	<p>Accession number PF00934</p> <p>Definition: PE family</p> <p>Author Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_253 (release 3.0)</p> <p>Gathering cutoffs: -20 -20</p> <p>Trusted cutoffs: -10.80 -10.80</p> <p>Noise cutoffs: -20.60 -20 60</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98295987</p> <p>Reference Title: Deciphering the biology of Mycobacterium tuberculosis from</p> <p>Reference Title: the complete genome sequence</p> <p>Reference Author: Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Reference Author: Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin</p> <p>Reference Author: N, Holroyd S, Hornsby T, Jagels K, Barrell BG. et al;</p> <p>Reference Location: Nature 1998;393:537-544.</p> <p>Database Reference: INTERPRO; IPR000084;</p> <p>Comment: This family named after a PE motif near to the amino terminus of the domain. The PE family of proteins all contain an amino-terminal region of about 110 amino acids. The carboxyl terminus of this family are variable and fall into several classes. The largest class of PE proteins is the highly repetitive PGRS class which have a high glycine content.</p> <p>Comment: The function of these proteins is uncertain but it has been suggested that they may be related to antigenic variation of Mycobacterium tuberculosis [1].</p> <p>Number of members: 90</p>
Pep_deformylase		Polypeptide deformylase	<p>Accession number: PF01327</p> <p>Definition: Polypeptide deformylase</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Sarah Teichmann</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 157.40 157 40</p> <p>Noise cutoffs: -29.00 -29.00</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97002011</p> <p>Reference Title: A new subclass of the zinc metalloproteases superfamily revealed by the solution structure of peptide deformylase.</p> <p>Reference Author: Meinnel T, Blanquet S, Dardel F;</p> <p>Reference Location: J Mol Biol 1996;262:375-386.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 98332750</p>

			<p>Reference Title: Solution structure of nickel-peptide deformylase.</p> <p>Reference Author: Dardel F, Ragusa S, Lazennec C, Blanquet S, Meinne T;</p> <p>Reference Location: J Mol Biol 1998;280:501-513.</p> <p>Database Reference: SCOP; 1def; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR000181,</p> <p>Database Reference: PDB; 2def; 4; 142;</p> <p>Database Reference: PDB; 1def; 4, 142;</p> <p>Database Reference: PDB; 1dff; 4, 142;</p> <p>Database Reference: PDB; 1bsj A; 4; 142;</p> <p>Database Reference: PDB; 1bsk A; 4; 142,</p> <p>Database Reference: PDB; 1bs4 A; 4; 142;</p> <p>Database Reference: PDB; 1bs4 B; 504; 642,</p> <p>Database Reference: PDB; 1bs4 C; 1004; 1142;</p> <p>Database Reference: PDB; 1bs5 A; 4; 142;</p> <p>Database Reference: PDB; 1bs5 B; 504, 642,</p> <p>Database Reference: PDB; 1bs5 C; 1004; 1142;</p> <p>Database Reference: PDB; 1bs6 A; 4; 142;</p> <p>Database Reference: PDB; 1bs6 B; 504; 642,</p> <p>Database Reference: PDB; 1bs6 C; 1004; 1142;</p> <p>Database Reference: PDB; 1bs7 A; 4; 142;</p> <p>Database Reference: PDB; 1bs7 B; 504; 642;</p> <p>Database Reference: PDB; 1bs7 C; 1004; 1142;</p> <p>Database Reference: PDB; 1bs8 A; 4; 142;</p> <p>Database Reference: PDB; 1bs8 B; 504; 642,</p> <p>Database Reference: PDB; 1bs8 C; 1004; 1142;</p> <p>Database Reference: PDB; 1bsz A; 4; 142,</p> <p>Database Reference: PDB; 1bsz B; 504, 642,</p> <p>Database Reference: PDB; 1bsz C; 1004, 1142;</p> <p>Database Reference: PDB; 1icj A; 4, 142;</p> <p>Database Reference: PDB; 1icj B; 504; 642;</p> <p>Database Reference: PDB; 1icj C; 1004; 1142;</p> <p>Database reference: PFAMB: PB041251;</p> <p>Number of members: 25</p>
Peptidase C 15		Pyroglutamyl peptidase	<p>Accession number. PF01470</p> <p>Definition: Pyroglutamyl peptidase</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw_manual</p> <p>Source of seed members: [1]</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs. 436.10 436 10</p> <p>Noise cutoffs: -155 40 -155.40</p> <p>HMM build command line. hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 99216536</p> <p>Reference Title: The crystal structure of pyroglutamyl peptidase I from</p> <p>Reference Title: bacillus amyloliquefaciens reveals a new structure for a</p> <p>Reference Title: cysteine protease.</p> <p>Reference Author: Odagaki Y, Hayashi A, Okada K, Hirotsu K, Kabashima T, Ito</p> <p>Reference Author: K, Yoshimoto T, Tsuru D, Sato M, Clardy J</p> <p>Reference Location: Structure 1999;7:399-411.</p> <p>Database Reference: SCOP, 1aug; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: MEROPS; C15;</p> <p>Database Reference: INTERPRO, IPR000816;</p> <p>Database Reference: PDB; 1a2z A; 2, 209;</p> <p>Database Reference: PDB; 1a2z B; 2; 209;</p> <p>Database Reference: PDB; 1a2z C; 2; 209;</p> <p>Database Reference: PDB; 1a2z D; 2; 209;</p> <p>Database Reference: PDB; 1aug A; 3; 204;</p> <p>Database Reference: PDB; 1aug B; 213; 414,</p> <p>Database Reference: PDB; 1aug C, 423; 624;</p> <p>Database Reference: PDB; 1aug D, 633; 834;</p> <p>Number of members: 10</p>
Peptidase_M 20	PDOC00613	ArgE / dapE / ACY1 / CPG2 / yscS family signatures	<p>The following enzymes have been shown [1,2,3] to be evolutionary and Functionally related</p> <p>- In the biosynthetic pathway from glutamate to arginine, the removal of an acetyl group from N2-acetylornithine can be catalyzed via two distinct enzymatic strategies depending on the organism. In some bacteria and in fungi, the acetyl group is transferred on glutamate by glutamate</p>

			<p>acetyltransferase (EC 2.3.1.35) while in enterobacteria such as <i>Escherichia coli</i>, it is hydrolyzed by acetylornithine deacetylase (EC 3.5.1.16) (acetylornithinase) (AO) (gene <i>argE</i>). AO is a homodimeric cobalt-dependent enzyme which displays broad specificity and can also deacylates substrates such as acetylarginine, acetylhistidine, acetylglutamate semialdehyde, etc.</p> <p>- Succinyldiaminopimelate desuccinylase (EC 3.5.1.18) (SDAP) (gene <i>dapE</i>) is the enzyme which catalyzes the fifth step in the biosynthesis of lysine from aspartate semialdehyde: the hydrolysis of succinyl-diaminopimelate to diaminopimelate and succinate. SDAP is an enzyme that requires cobalt or zinc as a cofactor</p> <p>- Aminoacylase-1 [4] (EC 3.5.1.14) (N-acyl-L-amino-acid amidohydrolase) (ACY1). ACY1 is a homodimeric zinc-binding mammalian enzyme that catalyzes the hydrolysis of N-alpha-acylated amino acids (except for aspartate)</p> <p>- Carboxypeptidase G2 (EC 3.4.17.11) (folate hydrolase G2) (gene <i>cpg2</i>) from <i>Pseudomonas</i> strain RS-16. This enzyme catalyzes the hydrolysis of reduced and non-reduced folates to pterates and glutamate G2 is a homodimeric zinc-dependent enzyme.</p> <p>- Vacuolar carboxypeptidase S (EC 3.4.17.4) (<i>yscS</i>) from yeast (gene <i>CPS1</i>)</p> <p>- Peptidase T (EC 3.4.11.-) (gene <i>pepT</i>) (tripeptidase) from bacteria. This enzyme catalyzes a variety of tripeptides containing N-terminal methionine, leucine, or phenylalanine</p> <p>- Xaa-His dipeptidase (EC 3.4.13.3) (carnosinase) from <i>Lactobacillus</i> (gene <i>pepV</i>) [5], a metalloenzyme with activity against beta-alanyl-dipeptides including carnosine (beta-alanyl-histidine)</p> <p>These enzymes share a few characteristics. They hydrolyse peptidic bonds in Substrates that share a common structure, they are dependent on cobalt or zinc For their activity and they are proteins of 40 Kd to 60 Kd with a number of Regions of sequence similarity.</p> <p>As signature patterns for these proteins, we selected two of the conserved Regions The first pattern contains a conserved histidine which could be Involved in binding metal ions and the second pattern contains a number of Conserved charged residues.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIV]-[GALMY]-[LIVMF]-x-[GSA]-H-x-D-[TV]-[STAV] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 6.</p> <p>Consensus pattern [GSTAI]-[SANQ]-D-x-K-[GSACN]-x(2)-[LIVMA]-x(2)-[LIVMFY]-x(14,17)-[LIVM]-x-[LIVMF]-[LIVMSTAG]-[LIVMFA]-x(2)-[DNG]-E-E-x-[GSTN] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note these proteins belong to families M20A/M20B in the classification of peptidases [6,E1]. Last update November 1997 / Patterns and text revised. References [1] Meinzel T., Schmitt E., Mechulam Y., Blanquet S. J. Bacteriol. 174:2323-2331(1992).</p> <p>[2] Boyen A., Charlier D., Sakanyan V., Mett I., Glansdorff N Gene 116:1-6(1992).</p> <p>[3] Miller C.G., Miller J.L., Bagga D.A J. Bacteriol. 173:3554-3558(1991).</p> <p>[4] Mitta M., Ohnogi H., Yamamoto A., Kato I., Sakiyama F., Tsunasawa S. J. Biochem. 112:737-742(1992).</p> <p>[5]</p>
--	--	--	--

			<p>Vongerichten K., Klein J., Matern H., Plapp R Microbiology 140:2591-2600(1994).</p> <p>[6] Rawlings N D., Barrett A J. Meth. Enzymol 248:183-228(1995)</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?peptidas.txt</p>
Peptidase_M 3	PDOC00129	Neutral zinc metallopeptidases, zinc-binding region signature	<p>The majority of zinc-dependent metallopeptidases (with the notable exception of the carboxypeptidases) share a common pattern of primary structure [1,2,3] in the part of their sequence involved in the binding of zinc, and can be grouped together as a superfamily, known as the metzincins, on the basis of this sequence similarity. They can be classified into a number of distinct families [4,E1] which are listed below along with the proteases which are currently known to belong to these families</p> <p>Family M1</p> <ul style="list-style-type: none"> - Bacterial aminopeptidase N (EC 3.4.11.2) (gene pepN) - Mammalian aminopeptidase N (EC 3.4.11.2) - Mammalian glutamyl aminopeptidase (EC 3.4.11.7) (aminopeptidase A) It may play a role in regulating growth and differentiation of early B-lineage cells. - Yeast aminopeptidase yscII (gene APE2). - Yeast alanine/arginine aminopeptidase (gene AAP1). - Yeast hypothetical protein YIL137c. - Leukotriene A-4 hydrolase (EC 3.3.2.6). This enzyme is responsible for the hydrolysis of an epoxide moiety of LTA-4 to form LTB-4. It has been shown that it binds zinc and is capable of peptidase activity <p>Family M2</p> <ul style="list-style-type: none"> - Angiotensin-converting enzyme (EC 3.4.15.1) (dipeptidyl carboxypeptidase I) (ACE) the enzyme responsible for hydrolyzing angiotensin I to angiotensin II. There are two forms of ACE: a testis-specific isozyme and a somatic isozyme which has two active centers. <p>Family M3</p> <ul style="list-style-type: none"> - Thimet oligopeptidase (EC 3.4.24.15). a mammalian enzyme involved in the cytoplasmic degradation of small peptides. - Neurolysin (EC 3.4.24.16) (also known as mitochondrial oligopeptidase M or microsomal endopeptidase). - Mitochondrial intermediate peptidase precursor (EC 3.4.24.59) (MIP) It is involved in the second stage of processing of some proteins imported in the mitochondrion. - Yeast saccharolysin (EC 3.4.24.37) (proteinase yscD). - Escherichia coli and related bacteria dipeptidyl carboxypeptidase (EC 3.4.15.5) (gene dcp). - Escherichia coli and related bacteria oligopeptidase A (EC 3.4.24.70) (gene opdA or prIC). - Yeast hypothetical protein YKL134c. <p>Family M4</p> <ul style="list-style-type: none"> - Thermostable thermolysins (EC 3.4.24.27), and related thermolabile neutral proteases (bacillolysins) (EC 3.4.24.28) from various species of Bacillus. - Pseudolysin (EC 3.4.24.26) from Pseudomonas aeruginosa (gene lasB). - Extracellular elastase from Staphylococcus epidermidis. - Extracellular protease prt1 from Erwinia carotovora. - Extracellular minor protease smp from Serratia marcescens. - Vibriolysin (EC 3.4.24.25) from various species of Vibrio. - Protease prtA from Listeria monocytogenes. - Extracellular proteinase proA from Legionella pneumophila. <p>Family M5</p> <ul style="list-style-type: none"> - Mycolysin (EC 3.4.24.31) from Streptomyces cacaoi. <p>Family M6</p> <ul style="list-style-type: none"> - Immune inhibitor A from Bacillus thuringiensis (gene ina). Ina degrades two classes of insect antibacterial proteins, attacins and cecropins. <p>Family M7</p> <ul style="list-style-type: none"> - Streptomyces extracellular small neutral proteases

			<p>Family M8</p> <ul style="list-style-type: none"> - Leishmanolysin (EC 3.4.24.36) (surface glycoprotein gp63). a cell surface protease from various species of <i>Leishmania</i> <p>Family M9</p> <ul style="list-style-type: none"> - Microbial collagenase (EC 3.4.24.3) from <i>Clostridium perfringens</i> and <i>Vibrio alginolyticus</i>.
			<p>Family M10A</p> <ul style="list-style-type: none"> - Serralysin (EC 3.4.24.40). an extracellular metalloprotease from <i>Serratia</i>. - Alkaline metalloproteinase from <i>Pseudomonas aeruginosa</i> (gene <i>aprA</i>). - Secreted proteases A, B, C and G from <i>Erwinia chrysanthemi</i>. - Yeast hypothetical protein YIL108w. <p>Family M10B</p> <ul style="list-style-type: none"> - Mammalian extracellular matrix metalloproteinases (known as matrixins) [5]: MMP-1 (EC 3.4.24.7) (interstitial collagenase), MMP-2 (EC 3.4.24.24) (72 Kd gelatinase), MMP-9 (EC 3.4.24.35) (92 Kd gelatinase), MMP-7 (EC 3.4.24.23) (matrylsin), MMP-8 (EC 3.4.24.34) (neutrophil collagenase), MMP-3 (EC 3.4.24.17) (stromelysin-1), MMP-10 (EC 3.4.24.22) (stromelysin-2), and MMP-11 (stromelysin-3). MMP-12 (EC 3.4.24.65) (macrophage metalloelastase). - Sea urchin hatching enzyme (envelysin) (EC 3.4.24.12) A protease that allows the embryo to digest the protective envelope derived from the egg extracellular matrix - Soybean metalloendoproteinase 1. <p>Family M11</p> <ul style="list-style-type: none"> - <i>Chlamydomonas reinhardtii</i> gamete lytic enzyme (GLE). <p>Family M12A</p> <ul style="list-style-type: none"> - Astacin (EC 3.4.24.21), a crayfish endoprotease - Meprin A (EC 3.4.24.18), a mammalian kidney and intestinal brush border metalloendopeptidase. - Bone morphogenic protein 1 (BMP-1), a protein which induces cartilage and bone formation and which expresses metalloendopeptidase activity. The <i>Drosophila</i> homolog of BMP-1 is the dorsal-ventral patterning protein <i>tolloid</i>. - Blastula protease 10 (BP10) from <i>Paracentrotus lividus</i> and the related protein SpAN from <i>Strongylocentrotus purpuratus</i>. - <i>Caenorhabditis elegans</i> protein <i>toh-2</i>. - <i>Caenorhabditis elegans</i> hypothetical protein F42A10.8. - Choriolytins L and H (EC 3.4.24.67) (also known as embryonic hatching proteins LCE and HCE) from the fish <i>Oryzias latipes</i>. These proteases participates in the breakdown of the egg envelope, which is derived from the egg extracellular matrix, at the time of hatching. <p>Family M12B</p> <ul style="list-style-type: none"> - Snake venom metalloproteinases [6]. This subfamily mostly groups proteases that act in hemorrhage. Examples are: adamalysin II (EC 3.4.24.46), atrolysin C/D (EC 3.4.24.42), atrolysin E (EC 3.4.24.44), fibrolase (EC 3.4.24.72), trimerelysin I (EC 3.4.25.52) and II (EC 3.4.25.53). - Mouse cell surface antigen MS2. <p>Family M13</p> <ul style="list-style-type: none"> - Mammalian neprilysin (EC 3.4.24.11) (neutral endopeptidase) (NEP). - Endothelin-converting enzyme 1 (EC 3.4.24.71) (ECE-1), which process the precursor of endothelin to release the active peptide. - Kell blood group glycoprotein, a major antigenic protein of erythrocytes. The Kell protein is very probably a zinc endopeptidase. - Peptidase O from <i>Lactococcus lactis</i> (gene <i>pepO</i>). <p>Family M27</p> <ul style="list-style-type: none"> - Clostridial neurotoxins, including tetanus toxin (TeTx) and the various botulinum toxins (BoNT). These toxins are zinc proteases that block neurotransmitter release by proteolytic cleavage of synaptic proteins such as synaptobrevins, syntaxin and SNAP-25 [7.8]. <p>Family M30</p> <ul style="list-style-type: none"> - <i>Staphylococcus hyicus</i> neutral metalloprotease. <p>Family M32</p>

950

		<p>- Thermostable carboxypeptidase 1 (EC 3.4.17.19) (carboxypeptidase Taq), an enzyme from <i>Thermus aquaticus</i> which is most active at high temperature.</p> <p>Family M34</p> <p>- Lethal factor (LF) from <i>Bacillus anthracis</i>, one of the three proteins composing the anthrax toxin</p> <p>Family M35</p> <p>- Deuterolysin (EC 3.4.24.39) from <i>Penicillium citrinum</i> and related proteases from various species of <i>Aspergillus</i>.</p> <p>Family M36</p> <p>- Extracellular elastinolytic metalloproteinases from <i>Aspergillus</i>.</p> <p>From the tertiary structure of thermolysin, the position of the residues acting as zinc ligands and those involved in the catalytic activity are known. Two of the zinc ligands are histidines which are very close together in the sequence: C-terminal to the first histidine is a glutamic acid residue which acts as a nucleophile and promotes the attack of a water molecule on the carbonyl carbon of the substrate. A signature pattern which includes the two histidine and the glutamic acid residues is sufficient to detect this superfamily of proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [GSTALIVN]-x(2)-H-E-[LIVMFYW]-{DEHRKP}-H-x-[LIVMFYWGSPQ] [The two H's are zinc ligands] [E is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for members of families M5, M7 and M11.</p> <p>Other sequence(s) detected in SWISS-PROT 57; including <i>Neurospora crassa</i> conidiation-specific protein 13 which could be a zinc-protease</p> <p>Last update July 1999 / Text revised.</p> <p>References</p> <p>[1] Jongeneel C V., Bouvier J., Bairoch A. FEBS Lett. 242 211-214(1989).</p> <p>[2] Murphy G.J.P., Murphy G., Reynolds J.J. FEBS Lett. 289 4-7(1991)</p> <p>[3] Bode W., Grams F., Reinemer P., Gomis-Rueth F.-X., Baumann U., McKay D.B., Stoecker W. Zoology 99:237-246(1996).</p> <p>[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).</p> <p>[5] Woessner J. Jr. FASEB J 5 2145-2154(1991).</p> <p>[6] Hite L.A., Fox J.W., Bjarnason J.B. Biol. Chem. Hoppe-Seyler 373.381-385(1992).</p> <p>[7] Montecucco C., Schiavo G. Trends Biochem. Sci. 18:324-327(1993)</p> <p>[8] Niemann H., Blasi J., Jahn R. Trends Cell Biol. 4.179-185(1994).</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?peptidas.txt</p>
Peptidase_M 48	Peptidase family M48	<p>Accession number: PF01435</p> <p>Definition: Peptidase family M48</p>

951

			<p>Author: Bateman A</p> <p>Alignment method of seed. Clustalw_manual</p> <p>Source of seed members. Swiss-Prot</p> <p>Gathering cutoffs: -35 -35</p> <p>Trusted cutoffs. -34.00 -34.00</p> <p>Noise cutoffs: -42.20 -42.20</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Database Reference MEROPS; M48;</p> <p>Database Reference INTERPRO; IPR001915;</p> <p>Database reference: PFAMB; PB008839;</p> <p>Database reference: PFAMB; PB041497;</p> <p>Number of members: 28</p>
Peptidase_S8	PDOC00125	Serine proteases, subtilase family, active sites	<p>Subtilases [1.2] are an extensive family of serine proteases whose catalytic activity is provided by a charge relay system similar to that of the trypsin family of serine proteases but which evolved by independent convergent evolution. The sequence around the residues involved in the catalytic triad (aspartic acid, serine and histidine) are completely different from that of the analogous residues in the trypsin serine proteases and can be used as signatures specific to that category of proteases.</p> <p>The subtilase family currently includes the following proteases:</p> <ul style="list-style-type: none"> - Subtilisins (EC 3.4.21.62), these alkaline proteases from various <i>Bacillus</i> species have been the target of numerous studies in the past thirty years. - Alkaline elastase YaB from <i>Bacillus</i> sp. (gene ale) - Alkaline serine exoprotease A from <i>Vibrio alginolyticus</i> (gene proA) - Aqualysin I from <i>Thermus aquaticus</i> (gene pstI). - AspA from <i>Aeromonas salmonicida</i> - Bacillopeptidase F (esterase) from <i>Bacillus subtilis</i> (gene bpf). - C5A peptidase from <i>Streptococcus pyogenes</i> (gene scpA). - Cell envelope-located proteases PI, PII, and PIII from <i>Lactococcus lactis</i> - Extracellular serine protease from <i>Serratia marcescens</i>. - Extracellular protease from <i>Xanthomonas campestris</i> - Intracellular serine protease (ISP) from various <i>Bacillus</i>. - Minor extracellular serine protease epr from <i>Bacillus subtilis</i> (gene epr). - Minor extracellular serine protease vpr from <i>Bacillus subtilis</i> (gene vpr). - Nisin leader peptide processing protease nisP from <i>Lactococcus lactis</i>. - Serotype-specific antigene 1 from <i>Pasteurella haemolytica</i> (gene ssa1). - Thermitase (EC 3.4.21.66) from <i>Thermoactinomyces vulgaris</i>. - Calcium-dependent protease from <i>Anabaena variabilis</i> (gene proA). - Halolysin from halophilic bacteria sp. 172p1 (gene hly). - Alkaline extracellular protease (AEP) from <i>Yarrowia lipolytica</i> (gene xpr2). - Alkaline proteinase from <i>Cephalosporium acremonium</i> (gene alp). - Cerevisin (EC 3.4.21.48) (vacuolar protease B) from yeast (gene PRB1) - Cuticle-degrading protease (pr1) from <i>Metarhizium anisopliae</i>. - KEX-1 protease from <i>Kluyveromyces lactis</i>. - Kexin (EC 3.4.21.61) from yeast (gene KEX-2) - Oryzin (EC 3.4.21.63) (alkaline proteinase) from <i>Aspergillus</i> (gene alp). - Proteinase K (EC 3.4.21.64) from <i>Tritirachium album</i> (gene proK). - Proteinase R from <i>Tritirachium album</i> (gene proR). - Proteinase T from <i>Tritirachium album</i> (gene proT). - Subtilisin-like protease III from yeast (gene YSP3). - Thermomycolin (EC 3.4.21.65) from <i>Malbranchea sulfurea</i>. - Furin (EC 3.4.21.85), neuroendocrine convertases 1 to 3 (NEC-1 to -3) and PACE4 protease from mammals, other vertebrates, and invertebrates. <p>These proteases are involved in the processing of hormone precursors at sites comprised of pairs of basic amino acid residues [3]</p> <ul style="list-style-type: none"> - Tripeptidyl-peptidase II (EC 3.4.14.10) (tripeptidyl aminopeptidase) from Human. - Prestalk-specific proteins tagB and tagC from slime mold [4]. Both proteins consist of two domains: a N-terminal subtilase catalytic domain and a C-terminal ABC transporter domain (see <PDOC00185>). <p>Description of pattern(s) and/or profile(s)</p>

952

			<p>Consensus pattern [STAI]-x-[LIVMF]-[LIVM]-D-[DSTA]-G-[LIVMFC]-x(2,3)-[DNH] [D is the active site residue] Sequences known to belong to this class detected by the pattern the majority of subtilases with a few exceptions. Other sequence(s) detected in SWISS-PROT 44.</p> <p>Consensus pattern H-G-[STM]-x-[VIC]-[STAGC]-[GS]-x-[LIVMA]-[STAGCLV]-[SAGM] [H is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for aspA and ssa1 which both seem to lack the histidine active site. Other sequence(s) detected in SWISS-PROT adenylate cyclase type VIII.</p> <p>Consensus pattern G-T-S-x-[SA]-x-P-x(2)-[STAVC]-[AG] [S is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for nisP, tagC and S marcescens extracellular serine protease. Other sequence(s) detected in SWISS-PROT 6.</p> <p>Note if a protein includes at least two of the three active site signatures, the probability of it being a serine protease from the subtilase family is 100%</p> <p>Note these proteins belong to family S8 in the classification of peptidases [5,E1]. Expert(s) to contact by email Brannigan J. jab5@vaxa.york.ac.uk</p> <p>Siezen R.J. siezen@nizo.nl</p> <p>Last update November 1997 / Patterns and text revised</p> <p>References [1] Siezen R.J., de Vos W.M., Leunissen J.A.M., Dijkstra B.W. Protein Eng. 4 719-737(1991).</p> <p>[2] Siezen R.J. (In) Proceeding subtilisin symposium, Hamburg, (1992).</p> <p>[3] Barr P.J. Cell 66:1-3(1991)</p> <p>[4] Shaulsky G., Kuspa A., Loomis W.F.: Genes Dev. 9:1111-1122(1995).</p> <p>[5] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994).</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?peptidas.txt</p>
Peptidase_S9	PDOC00587	Prolyl oligopeptidase family serine active site	<p>The prolyl oligopeptidase family [1,2,3] consist of a number of evolutionary related peptidases whose catalytic activity seems to be provided by a charge relay system similar to that of the trypsin family of serine proteases, but which evolved by independent convergent evolution. The known members of this family are listed below.</p> <ul style="list-style-type: none"> - Prolyl endopeptidase (EC 3.4.21.26) (PE) (also called post-proline cleaving enzyme). PE is an enzyme that cleaves peptide bonds on the C-terminal side of prolyl residues. The sequence of PE has been obtained from a mammalian species (pig) and from bacteria (Flavobacterium meningosepticum and Aeromonas hydrophila); there is a high degree of sequence conservation between these sequences - Escherichia coli protease II (EC 3.4.21.83) (oligopeptidase B) (gene prtB) which cleaves peptide bonds on the C-terminal side of lysyl and arginyl residues. - Dipeptidyl peptidase IV (EC 3.4.14.5) (DPP IV). DPP IV is an enzyme that removes N-terminal dipeptides sequentially from polypeptides having unsubstituted N-termini provided that the penultimate residue is proline - Yeast vacuolar dipeptidyl aminopeptidase A (DPAP A) (gene: STE13) which

953

			<p>IS</p> <p>responsible for the proteolytic maturation of the alpha-factor precursor.</p> <ul style="list-style-type: none"> - Yeast vacuolar dipeptidyl aminopeptidase B (DPAP B) (gene: DAP2). - Acylamino-acid-releasing enzyme (EC 3.4.19.1) (acyl-peptide hydrolase). <p>This enzyme catalyzes the hydrolysis of the amino-terminal peptide bond of an N-acetylated protein to generate a N-acetylated amino acid and a protein with a free amino-terminus.</p> <p>A conserved serine residue has experimentally been shown (in E.coli protease II as well as in pig and bacterial PE) to be necessary for the catalytic mechanism. This serine, which is part of the catalytic triad (Ser, His, Asp), is generally located about 150 residues away from the C-terminal extremity of these enzymes (which are all proteins that contains about 700 to 800 amino acids).</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern D-x(3)-A-x(3)-[LIVMFYW]-x(14)-G-x-S-x-G-G-[LIVMFYW](2) [S is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for yeast DPAP A</p> <p>Other sequence(s) detected in SWISS-PROT NONE</p> <p>Note these proteins belong to families S9A/S9B/S9C in the classification of peptidases [4,E1].</p> <p>Last update November 1997 / Text revised.</p> <p>References</p> <p>[1] Rawlings N D , Polgar L., Barrett A J. Biochem J. 279:907-911(1991).</p> <p>[2] Barrett A.J., Rawlings N.D Biol. Chem. Hoppe-Seyler 373:353-360(1992).</p> <p>[3] Polgar L., Szabo E. Biol. Chem. Hoppe-Seyler 373 361-366(1992).</p> <p>[4] Rawlings N.D., Barrett A.J. Meth. Enzymol. 244:19-61(1994)</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?peptidas.txt</p>
Peptidase_U 7		Peptidase family U7	<p>Accession number PF01343</p> <p>Definition: Peptidase family U7</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_707 (release 2.1)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs 47.60 47.60</p> <p>Noise cutoffs: -55.60 -55.60</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Database Reference MEROPS; U7;</p> <p>Database Reference INTERPRO; IPR002142;</p> <p>Number of members: 37</p>
PEP-utilizers	PDOC00527	PEP-utilizing enzymes signatures	<p>A number of enzymes that catalyze the transfer of a phosphoryl group from phosphoenolpyruvate (PEP) via a phospho-histidine intermediate have been shown to be structurally related [1,2,3,4] These enzymes are:</p> <ul style="list-style-type: none"> - Pyruvate,orthophosphate dikinase (EC 2.7.9.1) (PPDK). PPDK catalyzes the reversible phosphorylation of pyruvate and phosphate by ATP to PEP and diphosphate. In plants PPDK function in the direction of the formation of PEP, which is the primary acceptor of carbon dioxide in C4 and crassulacean acid metabolism plants. In some bacteria, such as Bacteroides symbiosus,

954

			<p>PPDK functions in the direction of ATP synthesis.</p> <ul style="list-style-type: none"> - Phosphoenolpyruvate synthase (EC 2.7.9.2) (pyruvate, water dikinase). This enzyme catalyzes the reversible phosphorylation of pyruvate by ATP to form PEP, AMP and phosphate, an essential step in gluconeogenesis when pyruvate and lactate are used as a carbon source - Phosphoenolpyruvate-protein phosphotransferase (EC 2.7.3.9) This is the first enzyme of the phosphoenolpyruvate-dependent sugar phosphotransferase system (PTS), a major carbohydrate transport system in bacteria. The PTS catalyzes the phosphorylation of incoming sugar substrates concomitant with their translocation across the cell membrane. The general mechanism of the PTS is the following: a phosphoryl group from PEP is transferred to enzyme-I (EI) of PTS which in turn transfers it to a phosphoryl carrier protein (HPr). Phospho-HPr then transfers the phosphoryl group to a sugar-specific permease <p>All these enzymes share the same catalytic mechanism. they bind PEP and transfer the phosphoryl group from it to a histidine residue. The sequence around that residue is highly conserved and can be used as a signature pattern for these enzymes. As a second signature pattern we selected a conserved region in the C-terminal part of the PEP-utilizing enzymes. The biological significance of this region is not yet known</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-[GA]-x-[STN]-x-H-[STA]-[STAV]-[LIVM](2)-[STAV]-[RG] [H is phosphorylated]</p> <p>Sequences known to belong to this class detected by the pattern ALL</p> <p>Other sequence(s) detected in SWISS-PROT NONE</p> <p>Consensus pattern [DEQSK]-x-[LIVMF]-S-[LIVMF]-G-[ST]-N-D-[LIVM]-x-Q-[LIVMFYGT]-[STALIV]-[LIVMFY]-[GAS]-x(2)-R</p> <p>Sequences known to belong to this class detected by the pattern ALL</p> <p>Other sequence(s) detected in SWISS-PROT NONE</p> <p>Last update December 1999 / Patterns and text revised.</p> <p>References</p> <p>[1] Reizer J., Hirschen C., Reizer A., Pham T.N., Saier M.H. Jr. Protein Sci. 2:506-521(1993).</p> <p>[2] Reizer J., Reizer A., Merrick M.J., Plunkett G. III, Rose D.J., Saier M.H. Jr. Gene 181:103-108(1996).</p> <p>[3] Pocalyko D.J., Carroll L.J., Martin B.M., Babbitt P.C., Dunaway-Mariano D. Biochemistry 29:10757-10765(1990).</p> <p>[4] Niersbach M., Kreuzaler F., Geerse R.H., Postma P., Hirsch H.J. Mol. Gen. Genet. 232:332-336(1992).</p>
PG_binding_2		Putative peptidoglycan binding domain	<p>Accession number: PF01476</p> <p>Definition: Putative peptidoglycan binding domain</p> <p>Author: Bateman A</p> <p>Alignment method of seed: HMM_built_from_alignment</p> <p>Source of seed members: Bateman A</p> <p>Gathering cutoffs: 22 22</p> <p>Trusted cutoffs: 22.40 22.10</p> <p>Noise cutoffs: 21 10 21.10</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 92324582</p> <p>Reference Title: Modular design of the Enterococcus hirae muramidase-2 and</p> <p>Reference Title: Streptococcus faecalis autolysin</p> <p>Reference Author: Joris B, Englebert S, Chu CP, Kariyama R, Daneo-Moore L,</p> <p>Reference Author: Shockman GD, Ghuysen JM;</p>

955

			<p>Reference Location: FEMS Microbiol Lett 1992;70:257-264.</p> <p>Database Reference: INTERPRO, IPR002482,</p> <p>Database reference: PFAMB; PB019287;</p> <p>Database reference: PFAMB; PB040847;</p> <p>Database reference: PFAMB; PB040977;</p> <p>Comment: This domain is about 40 residues long. It is found in a variety</p> <p>Comment: of enzymes involved in bacterial cell wall degradation [1].</p> <p>This</p> <p>Comment: domain may have a general peptidoglycan binding function.</p> <p>Number of members: 197</p>
phoslip	PDOC00109	Phospholipase A2 active sites signatures	<p>Phospholipase A2 (EC 3.1.1.4) (PA2) [1,2] is an enzyme which releases fatty acids from the second carbon group of glycerol. PA2's are small and rigid proteins of 120 amino-acid residues that have four to seven disulfide bonds. PA2 binds a calcium ion which is required for activity. The side chains of two conserved residues, a histidine and an aspartic acid, participate in a 'catalytic network'.</p> <p>Many PA2's have been sequenced from snakes, lizards bees and mammals. In the latter, there are at least four forms: pancreatic, membrane-associated as well as two less characterized forms. The venom of most snakes contains multiple forms of PA2. Some of them are presynaptic neurotoxins which inhibit neuromuscular transmission by blocking acetylcholine release from the nerve termini</p> <p>We derived two different signature patterns for PA2's. The first is centered on the active site histidine and contains three cysteines involved in disulfide bonds. The second is centered on the active site aspartic acid and also contains three cysteines involved in disulfide bonds.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern C-C-x(2)-H-x(2)-C [H is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL known functional PA2's. However, this pattern will not detect some snake toxins homologous with PA2 but which have lost their catalytic activity as well as otoconin-22, a Xenopus protein from the aragonitic otoconia which is also unlikely to be enzymatically active.</p> <p>Other sequence(s) detected in SWISS-PROT 15</p> <p>Consensus pattern [LIVMA]-C-[LIVMFYWPCST]-C-D-x(5)-C [D is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern the majority of functional and non-functional PA2's. Undetected sequences are bee PA2, gila monster PA2's, PA2 PL-X from habu and PA2 PA-5 from mulga.</p> <p>Other sequence(s) detected in SWISS-PROT 12.</p> <p>Expert(s) to contact by email</p> <p>Seilhamer J J. jeff@incyte.com</p> <p>Last update</p> <p>November 1995 / Patterns and text revised.</p> <p>References</p> <p>[1]</p> <p>Davidson F.F , Dennis E.A.</p> <p>J. Mol. Evol 31:228-238(1990)</p> <p>[2]</p> <p>Gomez F , Vandermeers A., Vandermeers-Piret M.-C , Herzog R., Rathe J., Stievenart M., Winand J., Christophe J.</p> <p>Eur. J Biochem. 186.23-33(1989).</p>
PI3_PI4_kinase	PDOC00710	Phosphatidylinositol 3- and 4-kinases signatures	<p>Phosphatidylinositol 3-kinase (PI3-kinase) (EC 2.7.1.137) [1] is an enzyme that phosphorylates phosphoinositides on the 3-hydroxyl group of the inositol ring. The exact function of the three products of PI3-kinase - PI-3-P, PI-3,4-P(2) and PI-3,4,5-P(3) - is not yet known, although it is proposed that they function as second messengers in cell signalling. Currently, three forms of PI3-kinase are known.</p> <p>- The mammalian enzyme which is a heterodimer of a 110 Kd catalytic chain</p>

2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 2679, 2680, 2681, 2682, 2683, 2684, 2685, 2686, 2687, 2688, 2689, 2690, 2691, 2692, 2693, 2694, 2695, 2696, 2697, 2698, 2699, 2700, 2701, 2702, 2703, 2704, 2705, 2706, 2707, 2708, 2709, 2710, 2711, 2712, 2713, 2714, 2715, 2716, 2717, 2718, 2719, 2720, 2721, 2722, 2723, 2724, 2725, 2726, 2727, 2728, 2729, 2730, 2731, 2732, 2733, 2734, 2735, 2736, 2737, 2738, 2739, 2740, 2741, 2742, 2743, 2744, 2745, 2746, 2747, 2748, 2749, 2750, 2751, 2752, 2753, 2754, 2755, 2756, 2757, 2758, 2759, 2760, 2761, 2762, 2763, 2764, 2765, 2766, 2767, 2768, 2769, 2770, 2771, 2772, 2773, 2774, 2775, 2776, 2777, 2778, 2779, 2780, 2781, 2782, 2783, 2784, 2785, 2786, 2787, 2788, 2789, 2790, 2791, 2792, 2793, 2794, 2795, 2796, 2797, 2798, 2799, 2800, 2801, 2802, 2803, 2804, 2805, 2806, 2807, 2808, 2809, 2810, 2811, 2812, 2813, 2814, 2815, 2816, 2817, 2818, 2819, 2820, 2821, 2822, 2823, 2824, 2825, 2826, 2827, 2828, 2829, 2830, 2831, 2832, 2833, 2834, 2835, 2836, 2837, 2838, 2839, 2840, 2841, 2842, 2843, 2844, 2845, 2846, 2847, 2848, 2849, 2850, 2851, 2852, 2853, 2854, 2855, 2856, 2857, 2858, 2859, 2860, 2861, 2862, 2863, 2864, 2865, 2866, 2867, 2868, 2869, 2870, 2871, 2872, 2873, 2874, 2875, 2876, 2877, 2878, 2879, 2880, 2881, 2882, 2883, 2884, 2885, 2886, 2887, 2888, 2889, 2890, 2891, 2892, 2893, 2894, 2895, 2896, 2897, 2898, 2899, 2900, 2901, 2902, 2903, 2904, 2905, 2906, 2907, 2908, 2909, 2910, 2911, 2912, 2913, 2914, 2915, 2916, 2917, 2918, 2919, 2920, 2921, 2922, 2923, 2924, 2925, 2926, 2927, 2928, 2929, 2930, 2931, 2932, 2933, 2934, 2935, 2936, 2937, 2938, 2939, 2940, 2941, 2942, 2943, 2944, 2945, 2946, 2947, 2948, 2949, 2950, 2951, 2952, 2953, 2954, 2955, 2956, 2957, 2958, 2959, 2960, 2961, 2962, 2963, 2964, 2965, 2966, 2967, 2968, 2969, 2970, 2971, 2972, 2973, 2974, 2975, 2976, 2977, 2978, 2979, 2980, 2981, 2982, 2983, 2984, 2985, 2986, 2987, 2988, 2989, 2990, 2991, 2992, 2993, 2994, 2995, 2996, 2997, 2998, 2999, 3000, 3001, 3002, 3003, 3004, 3005, 3006, 3007, 3008, 3009, 3010, 3011, 3012, 3013, 3014, 3015, 3016, 3017, 3018, 3019, 3020, 3021, 3022, 3023, 3024, 3025, 3026, 3027, 3028, 3029, 3030, 3031, 3032, 3033, 3034, 3035, 3036, 3037, 3038, 3039, 3040, 3041, 3042, 3043, 3044, 3045, 3046, 3047, 3048, 3049, 3050, 3051, 3052, 3053, 3054, 3055, 3056, 3057, 3058, 3059, 3060, 3061, 3062, 3063, 3064, 3065, 3066, 3067, 3068, 3069, 3070, 3071, 3072, 3073, 3074, 3075, 3076, 3077, 3078, 3079, 3080, 3081, 3082, 3083, 3084, 3085, 3086, 3087, 3088, 3089, 3090, 3091, 3092, 3093, 3094, 3095, 3096, 3097, 3098, 3099, 3100, 3101, 3102, 3103, 3104, 3105, 3106, 3107, 3108, 3109, 3110, 3111, 3112, 3113, 3114, 3115, 3116, 3117, 3118, 3119, 3120, 3121, 3122, 3123, 3124, 3125, 3126, 3127, 3128, 3129, 3130, 3131, 3132, 3133, 3134, 3135, 3136, 3137, 3138, 3139, 3140, 3141, 3142, 3143, 3144, 3145, 3146, 3147, 3148, 3149, 3150, 3151, 3152, 3153, 3154, 3155, 3156, 3157, 3158, 3159, 3160, 3161, 3162, 3163, 3164, 3165, 3166, 3167, 3168, 3169, 3170, 3171, 3172, 3173, 3174, 3175, 3176, 3177, 3178, 3179, 3180, 3181, 3182, 3183, 3184, 3185, 3186, 3187, 3188, 3189, 3190, 3191, 3192, 3193, 3194, 3195, 3196, 3197, 3198, 3199, 3200, 3201, 3202, 3203, 3204, 3205, 3206, 3207, 3208, 3209, 3210, 3211, 3212, 3213, 3214, 3215, 3216, 3217, 3218, 3219, 3220, 3221, 3222, 3223, 3224, 3225, 3226, 3227, 3228, 3229, 3230, 3231, 3232, 3233, 3234, 3235, 3236, 3237, 3238, 3239, 3240, 3241, 3242, 3243, 3244, 3245, 3246, 3247, 3248, 3249, 3250, 3251, 3252, 3253, 3254, 3255, 3256, 3257, 3258, 3259, 3260, 3261, 3262, 3263, 3264, 3265, 3266, 3267, 3268, 3269, 3270, 3271, 3272, 3273, 3274, 3275, 3276, 3277, 3278, 3279, 3280, 3281, 3282, 3283, 3284, 3285, 3286, 3287, 3288, 3289, 3290, 3291, 3292, 3293, 3294, 3295, 3296, 3297, 3298, 3299, 3300, 3301, 3302, 3303, 3304, 3305, 3306, 3307, 3308, 3309, 3310, 3311, 3312, 3313, 3314, 3315, 3316, 3317, 3318, 3319, 3320, 3321, 3322, 3323, 3324, 3325, 3326, 3327, 3328, 3329, 3330, 3331, 3332, 3333, 3334, 3335, 3336, 3337, 3338, 3339, 3340, 3341, 3342, 3343, 3344, 3345, 3346, 3347, 3348, 3349, 3350, 3351, 3352, 3353, 3354, 3355, 3356, 3357, 3358, 3359, 3360, 3361, 3362, 3363, 3364, 3365, 3366, 3367, 3368, 3369, 3370, 3371, 3372, 3373, 3374, 3375, 3376, 3377, 3378, 3379, 3380, 3381, 3382, 3383, 3384, 3385, 3386, 3387, 3388, 3389, 3390, 3391, 3392, 3393, 3394, 3395, 3396, 3397, 3398, 3399, 3400, 3401, 3402, 3403, 3404, 3405, 3406, 3407, 3408, 3409, 3410, 3411, 3412, 3413, 3414, 3415, 3416, 3417, 3418, 3419, 3420, 3421, 3422, 3423, 3424, 3425, 3426, 3427, 3428, 3429, 3430, 3431, 3432, 3433, 3434, 3435, 3436, 3437, 3438, 3439, 3440, 3441, 3442, 3443, 3444, 3445, 3446, 3447, 3448, 3449, 3450, 3451, 3452, 3453, 3454, 3455, 3456, 3457, 3458, 3459, 3460, 3461, 3462, 3463, 3464, 3465, 3466, 3467, 3468, 3469, 3470, 3471, 3472, 3473, 3474, 3475, 3476, 3477, 3478, 3479, 3480, 3481, 3482, 3483, 3484, 3485, 3486, 3487, 3488, 3489, 3490, 3491, 3492, 3493, 3494, 3495, 3496, 3497, 3498, 3499, 3500, 3501, 3502, 3503, 3504, 3505, 3506, 3507, 3508, 3509, 3510, 3511, 3512, 3513, 3514, 3515, 3516, 3517, 3518, 3519, 3520, 3521, 3522, 3523, 3524, 3525, 3526, 3527, 3528, 3529, 3530, 3531, 3532, 3533, 3534, 3535, 3536, 3537, 3538, 3539, 3540, 3541, 3542, 3543, 3544, 3545, 3546, 3547, 3548, 3549, 3550, 3551, 3552, 3553, 3554, 3555, 3556, 3557, 3558, 3559, 3560, 3561, 3562, 3563, 3564, 3565, 3566, 3567, 3568, 3569, 3570, 3571, 3572, 3573, 3574, 3575, 3576, 3577, 3578, 3579, 3580, 3581, 3582, 3583, 3584, 3585, 3586, 3587, 3588, 3589, 3590, 3591, 3592, 3593, 3594, 3595, 3596, 3597, 3598, 3599, 3600, 3601, 3602, 3603, 3604, 3605, 3606, 3607, 3608, 3609, 3610, 3611, 3612, 3613, 3614, 3615, 3616, 3617, 3618, 3619, 3620, 3621, 3622, 3623, 3624, 3625, 3626, 3627, 3628, 3629, 3630, 3631, 3632, 3633, 3634, 3635, 3636, 3637, 3638, 3639, 3640, 3641, 3642, 3643, 3644, 3645, 3646, 3647, 3648, 3649, 3650, 3651, 3652, 3653, 3654, 3655, 3656, 3657, 3658, 3659, 3660, 3661, 3662, 3663, 3664, 3665, 3666, 3667, 3668, 3669, 3670, 3671, 3672, 3673, 3674, 3675, 3676, 3677, 3678, 3679, 3680, 3681, 3682, 3683, 3684, 3685, 3686, 3687, 3688, 3689, 3690, 3691, 3692, 3693, 3694, 3695, 3696, 3697, 3698, 3699, 3700, 3701, 3702, 3703, 3704, 3705, 3706, 3707, 3708, 3709, 3710, 3711, 3712, 3713, 3714, 3715, 3716, 3717, 3718, 3719, 3720, 3721, 3722, 3723, 3724, 3725, 3726, 3727, 3728, 3729, 3730, 3731, 3732, 3733, 3734, 3735, 3736, 3737, 3738, 3739, 3740, 3741, 3742, 3743, 3744, 3745, 3746, 3747, 3748, 3749, 3750, 3751, 3752, 3753, 3754, 3755, 3756, 3757, 3758, 3759, 3760, 3761, 3762, 3763, 3764, 3765, 3766, 3767, 3768, 3769, 3770, 3771, 3772, 3773, 3774, 3775, 3776, 3777, 3778, 3779, 3780, 3781, 3782, 3783, 3784, 3785, 3786, 3787, 3788, 3789, 3790, 3791, 3792, 3793, 3794, 3795, 3796, 3797, 3798, 3799, 3800, 3801, 3802, 3803, 3804, 3805, 3806, 3807, 3808, 3809, 3810, 3811, 3812, 3813, 3814, 3815, 3816, 3817, 3818, 3819, 3820, 3821, 3822, 3823, 3824, 3825, 3826, 3827, 3828, 3829, 3830, 3831, 3832, 3833, 3834, 3835, 383

			<p>(p110) and an 85 Kd subunit (p85) which allows it to bind to activated tyrosine protein kinases. There are at least two different types of p100 subunits (alpha and beta)</p> <ul style="list-style-type: none"> - Yeast TOR1/DRR1 and TOR2/DRR2 [2], PI3-kinases required for cell cycle activation. Both are proteins of about 280 Kd. - Yeast VPS34 [3], a PI3-kinase involved in vacuolar sorting and segregation. VPS34 is a protein of about 100 Kd. - Arabidopsis thaliana and soybean VPS34 homologs. <p>Phosphatidylinositol 4-kinase (PI4-kinase) (EC 2.7.1.67) [4] is an enzyme that acts on phosphatidylinositol (PI) in the first committed step in the production of the second messenger inositol-1,4,5,-trisphosphate. Currently the following forms of PI4-kinases are known:</p> <ul style="list-style-type: none"> - Human PI4-kinase alpha. - Yeast PIK1, a nuclear protein of 120 Kd. - Yeast STT4, a protein of 214 Kd <p>The PI3- and PI4-kinases share a well conserved domain at their C-terminal section: this domain seems to be distantly related to the catalytic domain of protein kinases [2]. We developed two signature patterns from the best conserved parts of this domain.</p> <p>Four additional proteins belong to this family:</p> <ul style="list-style-type: none"> - Mammalian FKBP-rapamycin associated protein (FRAP) [5], which acts as the target for the cell-cycle arrest and immunosuppressive effects of the FKBP12-rapamycin complex. - Yeast protein ESR1 [6] which is required for cell growth, DNA repair and meiotic recombination - Yeast protein TEL1 which is involved in controlling telomere length. - Yeast hypothetical protein YHR099w, a distantly related member of this family. - Fission yeast hypothetical protein SpAC22E12 16C. <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVMFAC]-K-x(1.3)-[DEA]-[DE]-[LIVMC]-R-Q-[DE]-x(4)-Q Sequences known to belong to this class detected by the pattern ALL, except for yeast YHR099w Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [GS]-x-[AV]-x(3)-[LIVM]-x(2)-[FYH]-[LIVM](2)-x-[LIVMF]-x-D-R-H-x(2)-N Sequences known to belong to this class detected by the pattern ALL, except for yeast YHR099w Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Patterns and text revised.</p> <p>References [1] Hiles I.D., Otsu M., Volinia S., Fry M.J., Gout I., Dhand R., Panayotou G., Ruiz-Larrea F., Thompson A., Totty N.F., Hsuan J.J., Courtneidge S.A., Parker P.J., Waterfield M.D. Cell 70:419-429(1992).</p> <p>[2] Kunz J., Henriquez R., Schneider U., Deuter-Reinhard M., Movva N., Hall M.N. Cell 73:585-596(1993)</p> <p>[3] Schu P.V., Takegawa K., Fry M.J., Stack J.H., Waterfield M.D., Emr S.D. Science 260:88-91(1993).</p> <p>[4] Garcia-Bustos J.F., Marini F., Stevenson I., Frei C., Hall M.N. EMBO J 13:2352-2361(1994).</p> <p>[5] Brown E.J., Albers M.W., Shin T.B., Ichikawa K., Keith C.T., Lane W.S., Schreiber S.L.</p>
--	--	--	--

			<p>Nature 369:756-758(1994).</p> <p>[6] Kato R., Ogawa H. Nucleic Acids Res. 22:3104-3112(1994).</p>
P-II	PDOC00439	P-II protein signatures	<p>The P-II protein (gene glnB) is a bacterial protein important for the control of glutamine synthetase [1,2,3]. In nitrogen-limiting conditions, when the ratio of glutamine to 2-ketoglutarate decreases, P-II is uridylylated on a tyrosine residue to form P-II-UMP. P-II-UMP allows the deadenylation of glutamine synthetase (GS), thus activating the enzyme. Conversely, in nitrogen excess, P-II-UMP is deuridylylated and then promotes the adenylation of GS. P-II also indirectly controls the transcription of the GS gene (glnA) by preventing NR-II (ntrB) to phosphorylate NR-I (ntrC) which is the transcriptional activator of glnA. Once P-II is uridylylated, these events are reversed.</p> <p>P-II is a protein of about 110 amino acid residues extremely well conserved. The tyrosine which is uridylylated is located in the central part of the protein.</p> <p>In cyanobacteria, P-II seems to be phosphorylated on a serine residue rather than being uridylylated.</p> <p>In methanogenic archaeobacteria, the nitrogenase iron protein gene (nifH) is followed by two open reading frames highly similar to the eubacterial P-II protein [4]. These proteins could be involved in the regulation of nitrogen fixation.</p> <p>In the red alga, <i>Porphyra purpurea</i>, there is a glnB homolog encoded in the chloroplast genome.</p> <p>Other proteins highly similar to glnB are:</p> <ul style="list-style-type: none"> - <i>Bacillus subtilis</i> protein nrgB [5]. - <i>Escherichia coli</i> hypothetical protein ybaI [6] <p>We developed two signature patterns for P-II protein. The first one is a conserved stretch (in eubacteria) of six residues which contains the uridylylated tyrosine, the other is derived from a conserved region in the C-terminal part of the P-II protein.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern Y-[KR]-G-[AS]-[AE]-Y [The second Y is uridylylated] Sequences known to belong to this class detected by the pattern ALL glnB's from eubacteria. Other sequence(s) detected in SWISS-PROT 4.</p> <p>Consensus pattern [ST]-x(3)-G-[DY]-G-[KR]-[IV]-[FW]-[LIVM]-x(2)-[LIVM] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Patterns and text revised.</p> <p>References</p> <p>[1] Magasanik B. Brochimie 71:1005-1012(1989).</p> <p>[2] Holtel A., Merrick M Mol. Gen. Genet. 215:134-138(1988)</p> <p>[3] Cheah E., Carr P.D., Suffolk P.M., Vasuvedan S.G., Dixon N.E., Ollis D.L. Structure 2:981-990(1994).</p> <p>[4] Sibold L., Henriquet M., Possot O., Aubert J.-P. Res. Microbiol. 142:5-12(1991).</p> <p>[5] Wray L.V. Jr., Atkinson M.R., Fisher S.H. J. Bacteriol. 176:108-114(1994).</p>

			[6] Allikmets R., Gerrard B.C., Court D., Dean M.C. Gene 136 231-236(1993).
PLA2_B		Lysophospholipase catalytic domain	Accession number PF01735 Definition: Lysophospholipase catalytic domain Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_2127 (release 4 1) Gathering cutoffs: -283 -283 Trusted cutoffs: -185.70 -185.70 Noise cutoffs: -380.50 -380.50 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 94299545 Reference Title: Delineation of two functionally distinct domains of cytosolic phospholipase A2, a regulatory Ca(2+)-dependent Reference Title: lipid-binding domain and a Ca(2+)-independent catalytic domain. Reference Title: domain. Reference Author: Nalefski EA, Sultzman LA, Martin DM, Kriz RW, Towler PS, Reference Author: Knopf JL, Clark JD; Reference Location: J Biol Chem 1994;269:18239-18249. Reference Number: [2] Reference Medline: 94327513 Reference Title: The Saccharomyces cerevisiae PLB1 gene encodes a protein Reference Title: required for lysophospholipase and phospholipase B activity Reference Title: activity Reference Author: Lee KS, Patton JL, Fido M, Hines LK, Kohlwein SD, Paltauf Reference Author: F, Henry SA, Levin DE; Reference Location: J Biol Chem 1994;269:19725-19730 Database Reference: SCOP; 1rlw; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR002642; Database Reference: PDB; 1bci : 110; 138; Database Reference: PDB; 1c1y B, 1110, 1430, Database Reference: PDB; 1c1y A, 110; 498; Database Reference: PDB; 1rlw : 110; 140; Database Reference: PDB; 1c1y B; 1463; 1497, Database Reference: PDB; 1c1y B, 1539; 1717, Database Reference: PDB; 1c1y A; 539; 721; Comment: This family consists of Lysophospholipase / phospholipase B Comment: EC:3.1.1.5 and cytosolic phospholipase A2 EC:3.1.4 Comment: which also Comment: has a C2 domain C2. Comment: Phospholipase B enzymes catalyse the release of fatty acids from Comment: lysophospholipids and are capable in vitro of hydrolyzing all Comment: phospholipids extractable from yeast cells [1]. Comment: Cytosolic phospholipase A2 associates with natural membranes in Comment: response to physiological increases in Ca2+ and Comment: selectively Comment: hydrolyses arachidonyl phospholipids [2], the aligned region Comment: corresponds to the the carboxy-terminal Ca2+-independent catalytic Comment: domain of the protein as discussed in [2]. Number of members: 23
PLAT		PLAT/LH2 domain	Accession number PF01477 Definition: PLAT/LH2 domain Author: Bateman A Alignment method of seed: Manual Source of seed members: Bateman A Gathering cutoffs: 25 25 Trusted cutoffs: 29.40 29.40 Noise cutoffs: -7.90 -7.90 HMM build command line: hmmbuild HMM SEED

			<p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Database Reference: SCOP; 1lpa; fa: [SCOP-USA][CATH-PDBSUM]</p> <p>Database reference: PROSITE_PROFILE; PS50095,</p> <p>Database Reference: INTERPRO; IPR001024;</p> <p>Database Reference: PDB; 1lox ; 2; 112;</p> <p>Database Reference: PDB; 1hpl B; 336; 445;</p> <p>Database Reference: PDB; 1hpl A; 338; 447;</p> <p>Database Reference: PDB; 1eth C; 337; 403;</p> <p>Database Reference: PDB; 1eth A; 339; 405;</p> <p>Database Reference: PDB; 1eth C; 403; 445;</p> <p>Database Reference: PDB; 1eth A; 405; 447;</p> <p>Database Reference: PDB; 1rp1 ; 339; 449;</p> <p>Database Reference: PDB; 1bu8 A; 340; 407;</p> <p>Database Reference: PDB; 1bu8 A; 415; 452;</p> <p>Database Reference: PDB; 1gpl , 322; 334;</p> <p>Database Reference: PDB; 1ca1 ; 256; 370;</p> <p>Database Reference: PDB; 1qm6 A; 256; 370;</p> <p>Database Reference: PDB; 1qm6 B; 256; 370;</p> <p>Database Reference: PDB; 1qmd A; 256; 370;</p> <p>Database Reference: PDB; 1qmd B; 256; 370;</p> <p>Comment: This domain is found in a variety of membrane or lipid associated proteins It is called the PLAT</p> <p>Comment: (Polycystin-1, Lipoxxygenase, Alpha-Toxin) domain or</p> <p>Comment: LH2 (Lipoxxygenase homology) domain. The known structure</p> <p>Comment: of pancreatic lipase shows this domain binds to</p> <p>Comment: procolipase</p> <p>Comment: Colipase, which mediates membrane association.</p> <p>Comment: So it appears possible that this domain mediates</p> <p>Comment: membrane</p> <p>Comment: attachment via other protein binding partners The</p> <p>Comment: structure of this domain is known for many members of the</p> <p>Comment: family and is composed of a beta sandwich.</p> <p>Number of members: 82</p>
PLRV_ORF5	Potato leaf roll virus readthrough protein	<p>Accession number: PF01690</p> <p>Definition: Potato leaf roll virus readthrough protein</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1335 (release 4.1)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 116.40 116.40</p> <p>Noise cutoffs -285.50 -285.50</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 94233771</p> <p>Reference Title: Changes in the amino acid sequence of the coat protein readthrough domain of potato leafroll luteovirus affect the</p> <p>Reference Title: formation of an epitope and aphid transmission</p> <p>Reference Author: Jolly CA, Mayo MA;</p> <p>Reference Location: Virology 1994,201:182-185.</p> <p>Database Reference: INTERPRO; IPR002929;</p> <p>Comment: This family consists mainly of the potato leaf roll virus readthrough protein This is generated via a readthrough of open reading frame 3 a coat protein allowing</p> <p>Comment: transcription</p> <p>Comment: of open reading frame 5 to give an extended coat protein with a large c-terminal addition or read through domain [1].</p> <p>Comment: The readthrough protein is thought to play a role in the circulative aphid transmission of potato leaf roll virus [1].</p> <p>Comment: Also in the family is open reading frame 6 from beet western</p> <p>Comment: yellows virus and potato leaf roll virus both luteovirus and an unknown protein from cucurbit aphid-borne yellows</p> <p>Comment: virus a</p> <p>Comment: closterovirus.</p> <p>Number of members: 28</p>	
PMSR	Peptide methionine sulfoxide	<p>Accession number: PF01625</p> <p>Definition: Peptide methionine sulfoxide reductase</p> <p>Author: Bateman A</p>	

960

		reductase	<p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1111 (release 4.1)</p> <p>Gathering cutoffs: -62 -62</p> <p>Trusted cutoffs: -28.00 -28.00</p> <p>Noise cutoffs: -96.70 -96.70</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96353931</p> <p>Reference Title: Peptide methionine sulfoxide reductase contributes to the maintenance of adhesins in three major pathogens.</p> <p>Reference Author: Wizemann TM, Moskovitz J, Pearce BJ, Cundell D, Arvidson</p> <p>Reference Author: CG, So M, Weissbach H, Brot N, Masure HR,</p> <p>Reference Location: Proc Natl Acad Sci USA 1996;93:7985-7990.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 96312545</p> <p>Reference Title: Cloning the expression of a mammalian gene involved in the reduction of methionine sulfoxide residues in proteins.</p> <p>Reference Author: Moskovitz J, Weissbach H, Brot N;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1996;93:2095-2099.</p> <p>Database Reference: INTERPRO; IPR002569;</p> <p>Comment: This enzyme repairs damaged proteins. Methionine sulfoxide in proteins</p> <p>Comment: is reduced to methionine.</p> <p>Number of members: 28</p>
Pollen_allerg_2		Ribonuclease (pollen allergen)	<p>Accession number: PF01620</p> <p>Definition: Ribonuclease (pollen allergen)</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1050 (release 4.1)</p> <p>Gathering cutoffs: -3 -3</p> <p>Trusted cutoffs: 23 10 23.10</p> <p>Noise cutoffs: -29 40 -29.40</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 95246885</p> <p>Reference Title: Major allergen Phl p Vb in timothy grass is a novel pollen RNase.</p> <p>Reference Author: Bufe A, Schramm G, Keown MB, Schlaak M, Becker WM,</p> <p>Reference Location: Febs lett 1995;363:6-12</p> <p>Database Reference: INTERPRO; IPR002914;</p> <p>Database reference: PFAMB; PB037130;</p> <p>Comment: This family contains grass pollen proteins of group V.</p> <p>Comment: Swiss:Q40963 has been shown to possess ribonuclease activity [1].</p> <p>Number of members: 27</p>
POR_N		Pyruvate flavodoxin/ferredoxin oxidoreductase (N terminus)	<p>Accession number: PF01855</p> <p>Definition: Pyruvate flavodoxin/ferredoxin oxidoreductase (N terminus)</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_323 (release 4.2)</p> <p>Gathering cutoffs: -116 -116</p> <p>Trusted cutoffs: -113.60 -113.60</p> <p>Noise cutoffs: -119 50 -119.50</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96125254</p> <p>Reference Title: Molecular and phylogenetic characterization of pyruvate and 2-ketoisovalerate ferredoxin oxidoreductases from Pyrococcus furiosus and pyruvate ferredoxin oxidoreductase from Thermotoga maritima.</p> <p>Reference Author: Kletzin A, Adams MW,</p> <p>Reference Location: J Bacteriol 1996;178:248-257.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 94022264</p>

			<p>Reference Title. Growth of the cyanobacterium <i>Anabaena</i> on molecular nitrogen: NifJ is required when iron is limited.</p> <p>Reference Author: Bauer CC, Scappino L, Haselkorn R;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1993;90 8812-8816.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 99140300</p> <p>Reference Title: Crystal structures of the key anaerobic enzyme pyruvate:ferredoxin oxidoreductase, free and in complex with pyruvate.</p> <p>Reference Title: Chabriere E, Charon MH, Volbeda A, Pieulle L, Hatchikian</p> <p>Reference Author: EC, Fontecilla-Camps JC;</p> <p>Reference Location: Nat Struct Biol 1999;6:182-190.</p> <p>Database Reference: SCOP; 2pda; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: SCOP; 2pda; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR002880;</p> <p>Database Reference: PDB; 1b0p A; 43; 328,</p> <p>Database Reference: PDB; 1b0p B; 43; 328;</p> <p>Database Reference: PDB; 2pda A; 43; 328;</p> <p>Database Reference: PDB; 2pda B; 43; 328;</p> <p>Database reference: PFAMB: PB014847;</p> <p>Comment: This family includes the N terminal region of the pyruvate ferredoxin</p> <p>Comment: oxidoreductase, corresponding to the first two structural domains</p> <p>Comment: This region is involved in inter subunit contacts [3].</p> <p>Pyruvate</p> <p>Comment: oxidoreductase (POR) catalyses the final step in the fermentation</p> <p>Comment: of carbohydrates in anaerobic microorganisms [1]. This involves the</p> <p>Comment: oxidative decarboxylation of pyruvate with the participation of</p> <p>Comment: thiamine followed by the transfer of an acetyl moiety to coenzyme</p> <p>Comment: A for the synthesis of acetyl-CoA [1] The family also includes</p> <p>Comment: pyruvate flavodoxin oxidoreductase as encoded by the nifJ gene in</p> <p>Comment: cyanobacterium which is required for growth on molecular nitrogen</p> <p>Comment: when iron is limited [2]</p> <p>Number of members: 55</p>
PPE		PPE family	<p>Accession number: PF00823</p> <p>Definition: PPE family</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw_manual</p> <p>Source of seed members: Pfam-B_297 (release 3.0)</p> <p>Gathering cutoffs: -90 -90</p> <p>Trusted cutoffs: -88.20 -88.20</p> <p>Noise cutoffs: -105.30 -105.30</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98295987</p> <p>Reference Title: Deciphering the biology of <i>Mycobacterium tuberculosis</i> from</p> <p>Reference Title: the complete genome sequence.</p> <p>Reference Author:</p> <p>Reference Location: Nature 1998;393:537-544.</p> <p>Database Reference: INTERPRO; IPR000030;</p> <p>Database reference: PFAMB; PB040834;</p> <p>Comment: This family named after a PPE motif near to the amino terminus of the domain. The PPE family of proteins</p> <p>Comment: all contain an amino-terminal region of about 180 amino acids. The carboxyl terminus of this family</p> <p>Comment: are variable, and on the basis of this region fall into at least three groups. The MPTR subgroup has</p> <p>Comment: tandem copies of a motif NXGXGNXG. The second subgroup</p> <p>Comment: contains a conserved motif at about position 350</p> <p>Comment: The third group are only related in the amino terminal region</p> <p>Comment:</p>

			<p>Comment: The function of these proteins is uncertain but it</p> <p>Comment: has been suggested that they may be related to</p> <p>Comment: antigenic variation of <i>Mycobacterium tuberculosis</i> [1].</p> <p>Number of members: 75</p>
PRA-CH		Phosphoribosyl-AMP cyclohydrolase	<p>Accession number: PF01502</p> <p>Definition: Phosphoribosyl-AMP cyclohydrolase</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_782 (release 4.0)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 88.20 88.20</p> <p>Noise cutoffs: -44.30 -44.30</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 99129952</p> <p>Reference Title: N1-(5'-phosphoribosyl)adenosine-5'-monophosphate cyclohydrolase: purification and characterization of a unique metalloenzyme.</p> <p>Reference Author: D'Ordine RL, Klem TJ, Davisson VJ,</p> <p>Reference Location: Biochemistry 1999;38:1537-1546</p> <p>Database Reference: INTERPRO: IPR002496;</p> <p>Comment: This enzyme catalyses the third step in the histidine biosynthetic pathway. It requires Zn ions for activity.</p> <p>Number of members: 28</p>
PRA-PH		Phosphoribosyl-ATP pyrophosphohydrolase	<p>Accession number: PF01503</p> <p>Definition: Phosphoribosyl-ATP pyrophosphohydrolase</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_784 (release 4.0)</p> <p>Gathering cutoffs: 6 6</p> <p>Trusted cutoffs: 12 10 12 10</p> <p>Noise cutoffs: 1.00 1.00</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 79216449</p> <p>Reference Title: The product of the <i>his4</i> gene cluster in <i>Saccharomyces cerevisiae</i>. A trifunctional polypeptide</p> <p>Reference Author: Keesey JK Jr, Bigelis R, Fink GR;</p> <p>Reference Location: J Biol Chem 1979 Aug 10;254:7427-7433.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 86310274</p> <p>Reference Title: Primary and secondary structural homologies between the</p> <p>Reference Title: HIS4 gene product of <i>Saccharomyces cerevisiae</i> and the</p> <p>Reference Title: and <i>hisD</i> gene products of <i>Escherichia coli</i> and <i>Salmonella typhimurium</i></p> <p>Reference Author: Bruni CB, Carlomagno MS, Formisano S, Paoletti G;</p> <p>Reference Location: Mol Gen Genet 1986;203:389-396.</p> <p>Database Reference: INTERPRO: IPR002497;</p> <p>Comment: This enzyme catalyses the second step in the histidine biosynthetic pathway</p> <p>Number of members: 32</p>
PseudoU_synth_1		tRNA pseudouridine synthase	<p>Accession number: PF01416</p> <p>Definition: tRNA pseudouridine synthase</p> <p>Previous Pfam IDs: PseudoU_synth;</p> <p>Author: Howe K</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: swissprot</p> <p>Gathering cutoffs: 30 30</p> <p>Trusted cutoffs: 39.10 39.10</p> <p>Noise cutoffs: -55.00 -55.00</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98254513</p> <p>Reference Title: Transfer RNA-pseudouridine synthetase Pus1 of <i>Saccharomyces</i></p>

963

			<p>Reference Title: cerevisiae contains one atom of zinc essential for its native conformation and tRNA recognition.</p> <p>Reference Author: Arluisson V, Hountondji C, Robert B, Grosjean H,</p> <p>Reference Location: Biochemistry 1998;37:7268-7276.</p> <p>Database Reference: INTERPRO; IPR001406;</p> <p>Database reference: PFAMB; PB027500;</p> <p>Comment: Involved in the formation of pseudouridine at the anticodon stem</p> <p>Comment: and loop of transfer-RNAs</p> <p>Comment: Pseudouridine is an isomer of uridine (5-(beta-D-ribofuranosyl))</p> <p>Comment: uracil, and is the most abundant modified nucleoside found in</p> <p>Comment: all cellular RNAs.</p> <p>Comment: The TruA-like proteins also exhibit a conserved sequence with</p> <p>Comment: a strictly conserved aspartic acid, likely involved in catalysis</p> <p>Number of members: 31</p>
PseudoU_syn th_2		RNA pseudouridylate synthase	<p>Accession number: PF00849</p> <p>Definition: RNA pseudouridylate synthase</p> <p>Previous Pfam IDs: YABO;</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_421 (release 3.0)</p> <p>Gathering cutoffs: 20 20</p> <p>Trusted cutoffs: 20 90 20.90</p> <p>Noise cutoffs: -44 40 -44.40</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96079974</p> <p>Reference Title: A dual-specificity pseudouridine synthase from Escherichia coli synthase purified and cloned on the basis of its specificity for psi 746 in 23S RNA is also specific for psi 32 in tRNA(phe).</p> <p>Reference Author: Wrzesinski J, Nurse K, Bakin A, Lane BG, Ofengand J;</p> <p>Reference Location: RNA 1995;1:437-448.</p> <p>Database Reference: PROSITE, PDOC00869</p> <p>Database Reference: PROSITE, PDOC00885</p> <p>Database Reference: INTERPRO, IPR000613,</p> <p>Database reference: PFAMB; PB041160;</p> <p>Database reference: PFAMB; PB041232;</p> <p>Comment: Members of this family are involved in modifying bases in RNA molecules.</p> <p>Comment: They carry out the conversion of uracil bases to pseudouridine. This family</p> <p>Comment: includes RluD Swiss:P33643, a pseudouridylate synthase that converts</p> <p>Comment: specific uracils to pseudouridine in 23S rRNA. RluA from E. coli</p> <p>Comment: converts bases in both rRNA and tRNA [1].</p> <p>Number of members: 78</p>
PWI		PWI domain	<p>Accession number: PF01480</p> <p>Definition: PWI domain</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw_manual</p> <p>Source of seed members: [1]</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 64.40 64.40</p> <p>Noise cutoffs: -3.50 -3.50</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 10322432</p> <p>Reference Title: The PWI motif, a new protein domain in splicing factors</p> <p>Reference Author: Blencowe BJ, Ouzounis CA;</p> <p>Reference Location: Trends Biochem Sci 1999;24:179-180.</p> <p>Database Reference: INTERPRO; IPR002483;</p> <p>Number of members: 11</p>
R3H		R3H domain	<p>Accession number: PF01424</p>

964

			<p>R3H domain</p> <p>Definition: R3H domain</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Manual</p> <p>Source of seed members: Medline 99003905</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 59.30 59.30</p> <p>Noise cutoffs: 5.10 5.10</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 99003905</p> <p>Reference Title: The R3H motif: a domain that binds single-stranded nucleic acids.</p> <p>Reference Author: Grishin NV;</p> <p>Reference Location: Trends Biochem Sci 1998;23 329-330</p> <p>Database Reference: INTERPRO: IPR001374,</p> <p>Database reference: PFAM: PB041444;</p> <p>Comment: The name of the R3H domain comes from the characteristic spacing</p> <p>Comment: of the most conserved arginine and histidine residues</p> <p>The</p> <p>Comment: function of the domain is predicted to be binding ssDNA.</p> <p>Number of members: 28</p>
RepB_protein		Initiator RepB protein	<p>Accession number: PF01051</p> <p>Definition: Initiator RepB protein</p> <p>Author: Finn RD, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_313 (release 3 0)</p> <p>Gathering cutoffs: 14 14</p> <p>Trusted cutoffs: 19.00 16 20</p> <p>Noise cutoffs: 11.80 12.90</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98284148</p> <p>Reference Title: Replication and control of circular bacterial plasmids.</p> <p>Reference Author: del Solar G, Giraldo R, Ruiz-Echevarria MJ, Espinosa M,</p> <p>Reference Author: Diaz-Orejas R,</p> <p>Reference Location: Microbiol Mol Biol Rev 1998;62 434-464.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97324207</p> <p>Reference Title: Initiation of replication of plasmid pMV158: mechanisms of</p> <p>Reference Title: DNA strand-transfer reactions mediated by the initiator</p> <p>Reference Title: RepB protein.</p> <p>Reference Author: Moscoso M, Eritja R, Espinosa M;</p> <p>Reference Location: J Mol Biol 1997;268 840-856.</p> <p>Database Reference: INTERPRO: IPR000525;</p> <p>Database Reference: PDB; 1rep C, 198; 240;</p> <p>Database reference: PFAM; PB000509;</p> <p>Comment: This protein is an initiator of plasmid replication.</p> <p>Comment: RepB possesses nicking-closing (topoisomerase I) like activity.</p> <p>Comment: It is also able to perform a strand transfer reaction on ssDNA</p> <p>Comment: that contains its target.</p> <p>Number of members: 51</p>
Rhomboid		Rhomboid family	<p>Accession number: PF01694</p> <p>Definition: Rhomboid family</p> <p>Author: Schramm M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1399 (release 4.1)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 143.60 143.60</p> <p>Noise cutoffs: -43.60 -43.60</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 90249726</p> <p>Reference Title: rhomboid, a gene required for dorsoventral axis</p>

			<p>Reference Title establishment and peripheral nervous system development in</p> <p>Reference Title: <i>Drosophila melanogaster</i>.</p> <p>Reference Author: Bier E, Jan LY, Jan YN;</p> <p>Reference Location: Genes Dev 1990,4.190-203.</p> <p>Database Reference INTERPRO; IPR002610;</p> <p>Database reference: PFAMB; PB041113;</p> <p>Comment: This family contains integral membrane proteins that are related to <i>Drosophila</i> rhomboid protein Swiss:P20350.</p> <p>Members</p> <p>Comment: of this family are found in bacteria and eukaryotes. These proteins contain three strongly conserved histidines in the putative transmembrane regions that may be involved in the</p> <p>Comment: as yet unknown function of these proteins.</p> <p>Number of members: 27</p>
Ribosomal L 18ae		Ribosomal L18ae protein family	<p>Accession number. PF01775</p> <p>Definition: Ribosomal L18ae protein family</p> <p>Author: Bateman A</p> <p>Alignment method of seed. Clustalw</p> <p>Source of seed members: PSI-BLAST Q02543</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs 136.70 136.70</p> <p>Noise cutoffs: -99.80 -99.80</p> <p>HMM build command line hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Database Reference INTERPRO; IPR002670;</p> <p>Number of members: 11</p>
Ribosomal L 21p	PDOC00899	Ribosomal protein L21 signature	<p>Ribosomal protein L21 is one of the proteins from the large ribosomal subunit in <i>Escherichia coli</i>. L21 is known to bind to the 23S rRNA in the presence of L20. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities, groups:</p> <ul style="list-style-type: none"> - Eubacterial L21. - <i>Marchantia polymorpha</i> chloroplast L21. - Cyanelle L21. - Spinach chloroplast L21 (nuclear-encoded) <p>Eubacterial L21 is a protein of about 100 amino-acid residues, the mature form of the spinach chloroplast L21 has 200 residues. As a signature pattern, we selected a conserved region located in the C-terminal section of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [IVT]-x(3)-[KR]-x(3)-[KRQ]-K-x(6)-G-[HF]-R-[RQ]-x(2)-[ST]</p> <p>Sequences known to belong to this class detected by the pattern ALL</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update</p> <p>July 1999 / Pattern and text revised.</p>
Ribosomal L 22e		Ribosomal L22e protein family	<p>Accession number: PF01776</p> <p>Definition: Ribosomal L22e protein family</p> <p>Author: Bateman A</p> <p>Alignment method of seed. Clustalw</p> <p>Source of seed members. PSI-BLAST P56628</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 262.80 262.80</p> <p>Noise cutoffs: -52.00 -52.00</p> <p>HMM build command line hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Database Reference INTERPRO; IPR002671;</p> <p>Number of members: 11</p>
Ribosomal L 27e		Ribosomal L27e protein family	<p>Accession number: PF01777</p> <p>Definition: Ribosomal L27e protein family</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members. PSI-BLAST P51419</p> <p>Gathering cutoffs: 25 25</p>

966

			<p>Trusted cutoffs 326.90 326.90 Noise cutoffs: -47.80 -47.80 HMM build command line hmmbuild HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Database Reference INTERPRO; IPR001141; Number of members: 9</p>
Ribosomal_L 29	PDOC00501	Ribosomal protein L29 signature	<p>Ribosomal protein L29 is one of the proteins from the large ribosomal subunit. L29 belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups</p> <ul style="list-style-type: none"> - Eubacterial L29 - Red algal L29. - Archaeobacterial L29 - Mammalian L35 - Caenorhabditis elegans L35 (ZK652.4). - Yeast L35. <p>L29 is a protein of 63 to 138 amino-acid residues. As a signature pattern, we selected a conserved region located in the central section of L29.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [KNQS]-[PSTLN]-x(2)-[LIMFA]-[KRGSAN]-x-[LIVYSTA]-[KR]-[KRHQS]-[DESTANRL]-[LIV]-A-[KRCQVT]-[LIVMA] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT 2. Last update December 1999 / Pattern and text revised. References [1] Otaka E., Hashimoto T., Mizuta K. Protein Seq Data Anal. 5:285-300(1993).</p>
Ribosomal_L 31e	PDOC00881	Ribosomal protein L31e signature	<p>A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:</p> <ul style="list-style-type: none"> - Mammalian L31 [1] - Chlamydomonas reinhardtii L31. - Yeast L34. - Halobacterium marismortui HL30 [2]. <p>These proteins have 87 to 128 amino-acid residues. As a signature pattern, we selected a conserved region located in the central section.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern V-[KR]-[LIVM]-x(3)-[LIVM]-N-x-[AKH]-x-W-x-[KR]-G Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update July 1999 / Pattern and text revised. References [1] Tanaka T., Kuwano Y., Kuzumaki T., Ishikawa K., Ogata K. Eur. J. Biochem 162:45-48(1987).</p> <p>[2] Bergmann U., Arndt E. Biochim. Biophys Acta 1050:56-60(1990)</p>
Ribosomal_L 35Ae	PDOC00849	Ribosomal protein L35Ae signature	<p>A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:</p> <ul style="list-style-type: none"> - Vertebrate L35A - Caenorhabditis elegans L35A (F10E7.7). - Yeast L37A/L37B (Rp47) - Pyrococcus woesei L35A homolog [1].

			<p>These proteins have 87 to 110 amino-acid residues. As a signature pattern, we selected a highly conserved stretch of 22 residues in the C-terminal part of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-K-[LIVM]-x-R-x-H-G-x(2)-G-x-V-x-A-x-F-x(3)-[LI]-P Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE. Last update November 1997 / Pattern and text revised References [1] Ouzounis C , Kyriakides N., Sander C. Nucleic Acids Res. 23 565-570(1995).</p>
Ribosomal_L 35p	PDOC00721	Ribosomal protein L35 signature	<p>Ribosomal protein L35 is one of the proteins from the large subunit of the ribosome. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups:</p> <ul style="list-style-type: none"> - Eubacterial L35. - Plant chloroplast L35 (nuclear-encoded). - Red algal chloroplast L35. - Cyanelle L35. <p>L35 is a basic protein of 60 to 70 amino-acid residues. As a signature pattern we selected a conserved region in the N-terminal section.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVM]-K-[TV]-x(2)-[GSA]-[SAILV]-x-K-R-[LIVMFY]-[KRLS] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE Last update December 1999 / Pattern and text revised. References [1] Otaka E., Hashimoto T , Mizuta K. Protein Seq. Data Anal. 5 285-300(1993).</p>
Ribosomal_L 36e	PDOC00916	Ribosomal protein L36e signature	<p>A number of eukaryotic ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:</p> <ul style="list-style-type: none"> - Mammalian L36 [1] - Drosophila L36 (M(1)1B). - Caenorhabditis elegans L36 (F37C12.4). - Candida albicans L39. - Yeast YL39. <p>These proteins have 99 to 104 amino acids. As a signature pattern, we selected a conserved region in the central part of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern P-Y-E-[KR]-R-x-[LIVM]-[DE]-[LIVM](2)-[KR] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE. Last update November 1997 / First entry References [1] Chan Y.-L., Paz V., Olvera J., Wool I.G. Biochem. Biophys. Res. Commun. 192:849-853(1993).</p>
Ribosomal_L 37ae		Ribosomal L37ae protein family	<p>Accession number: PF01780 Definition: Ribosomal L37ae protein family Author: Bateman A</p>

968

			<p>Alignment method of seed: Clustalw Source of seed members: PSI-BLAST P54051 Gathering cutoffs: 25 25 Trusted cutoffs: 145.10 145 10 Noise cutoffs: -46.90 -46 90 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Database Reference INTERPRO; IPR002674; Comment: This ribosomal protein is found in archaeobacteria and Comment: eukaryotes. It contains four conserved cysteine Comment: residues that may bind to zinc. Number of members: 15</p>
Ribosomal_L 37e	PDOC00827	Ribosomal protein L37e signature	<p>A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:</p> <ul style="list-style-type: none"> - Mammalian L37 [1]. - Leishmania infantum L37 [2]. - Fission yeast YL35 [3]. - Halobacterium marismortui L37e (L35e) [4]. <p>These proteins have 56 to 96 amino-acid residues. As a signature pattern, we selected a highly conserved region located in the N-terminal part of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-T-x-[SA]-x-G-x-[KR]-x(3)-[STLR]-x(0,1)-H-x(2)-C-x-R-C-G Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update July 1999 / Pattern and text revised. References [1] Chan Y.-L., Paz V., Olvera J., Wool I.G Biochem. Biophys. Res. Commun 192:590-596(1993)</p> <p>[2] Myler P.J , Tripp C.A., Thomas L., Venkataraman G M., Merlin G , Stuart K Mol. Biochem. Parasitol. 62:147-152(1993).</p> <p>[3] Otaka E . Higo K -I., Itoh T. Mol. Gen. Genet. 191:519-524(1983).</p> <p>[4] Bergmann U., Wittmann-Liebold B Biochim. Biophys. Acta 1173:195-200(1993).</p>
Ribosomal_L 38e		Ribosomal L38e protein family	<p>Accession number PF01781 Definition: Ribosomal L38e protein family Author: Bateman A Alignment method of seed: Clustalw Source of seed members: PSI-BLAST P23411 Gathering cutoffs: 25 25 Trusted cutoffs: 127 60 127.60 Noise cutoffs: -24.50 -24.50 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 91207349 Reference Title: The primary structure of rat ribosomal protein L38. Reference Author: Kuwano Y, Olvera J, Wool IG, Reference Location: Biochem Biophys Res Commun 1991;175:551-555. Database Reference: INTERPRO, IPR002675; Number of members: 8</p>
Ribosomal_L 39	PDOC00050	Ribosomal protein L39e signature	<p>A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of:</p>

969

			<ul style="list-style-type: none"> - Mammalian L39 [1]. - Plants L39. - Yeast L46 [2]. - Archebacterial L39e [3]. <p>These proteins are very basic. About 50 residues long, they are the smallest proteins of eukaryotic-type ribosomes. As a signature pattern, we selected a conserved region in the C-terminal section of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [KRA]-T-x(3)-[LIVM]-[KRQF]-x-[NHS]-x(3)-R-[NHY]-W-R-R</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE</p> <p>Last update July 1998 / Pattern and text revised.</p> <p>References [1] Lin A., McNally J., Wool I G. J. Biol. Chem. 259:487-490(1984).</p> <p>[2] Leer R.J., van Raamsdonk-Duin M.M.C., Kraakman P. Mager W H , Planta R.J. Nucleic Acids Res. 13:701-709(1985)</p> <p>[3] Ramirez C , Louie K A , Matheson A T. FEBS Lett. 250 416-418(1989).</p>
Ribosomal_L 4	PDOC00724	Ribosomal protein L1e signature	<p>A number of eukaryotic and archaebacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists [1,2,3, 4] of:</p> <ul style="list-style-type: none"> - Vertebrate L1 (L4). - Drosophila L1 - Plant L1. - Yeast L2 (Rp2). - Fission yeast L2. - Halobacterium marismortui HmaL4 (HL6). - Methanococcus jannaschii MJ0177. <p>These proteins have 246 (archaeobacteria) to 427 (human) amino acids. As a signature pattern, we selected a conserved region in the N-terminal part of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern N-x(3)-[KRM]-x(2)-A-[LIVT]-x-S-A-[LIV]-x-A-[ST]-[SGA]-x(7)-[RK]-[GS]-H</p> <p>Sequences known to belong to this class detected by the pattern ALL</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update November 1997 / Pattern and text revised</p> <p>References [1] Rafti F , Gargiulo G., Manzi A , Malva C., Graziani F. Nucleic Acids Res. 17:456-456(1989).</p> <p>[2] Presutti C., Villa T., Bozzoni I. Nucleic Acids Res. 21:3900-3900(1993).</p> <p>[3] Bagni C., Mariottini P., Annesi F., Amaldi F. Arndt E., Kroemer W., Hatakeyama T Biochim. Biophys. Acta 1216:475-478(1993). J. Biol. Chem. 265:3034-3039(1990).</p>
Ribosomal S		Ribosomal	Accession number: PF01649

970

20p		protein S20	<p>Definition: Ribosomal protein S20</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1685 (release 4.1)</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 57 30 57 30</p> <p>Noise cutoffs: -25.50 -25.50</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 88230452</p> <p>Reference Title: Interaction of proteins S16, S17 and S20 with 16 S</p> <p>Reference Title: ribosomal RNA.</p> <p>Reference Author: Stern S, Changchien LM, Craven GR, Noller HF,</p> <p>Reference Location: J Mol Biol 1988;200:291-299.</p> <p>Database Reference: INTERPRO, IPR002583,</p> <p>Comment: Bacterial ribosomal protein S20 interacts with 16S rRNA [1]</p> <p>Number of members: 29</p>
Ribosomal_S 27e	PDOC00898	Ribosomal protein S27e signature	<p>A number of eukaryotic and archaeobacterial ribosomal proteins can be grouped on the basis of sequence similarities. One of these families consists of [1]</p> <ul style="list-style-type: none"> - Mammalian S27 (human S27 was originally known as metallopan-stimulin 1) - Chlamydomonas reinhardtii S27. - Entamoeba histolytica S27 - Yeast S27 - Archaeobacterial S27e. <p>These proteins have from 62 to 87 amino acids. They contain, in their central section, a putative zinc-finger region of the type C-x(2)-C-x(14)-C-x(2)-C. We have selected that region as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [QKT]-C-x(2)-C-x(6)-F-[GSD]-x-[PSA]-x(5)-C-x(2)-C-[GSA]-x(2)-[LV]-x(2)-P-x-G [The four C's are potential zinc ligands]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE</p> <p>Last update December 1999 / Pattern and text revised.</p> <p>References [1] Chan Y -L , Suzuki K., Olvera J., Wool I.G. Nucleic Acids Res. 21:649-655(1993).</p>
Ribosomal_S 3_C	PDOC00474	Ribosomal protein S3 signature	<p>Ribosomal protein S3 is one of the proteins from the small ribosomal subunit. In Escherichia coli, S3 is known to be involved in the binding of initiator Met-tRNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups:</p> <ul style="list-style-type: none"> - Eubacterial S3. - Algal and plant chloroplast S3. - Cyanelle S3. - Archaeobacterial S3. - Plant mitochondrial S3. - Vertebrate S3. - Insect S3. - Caenorhabditis elegans S3 (C23G10.3). - Yeast S3 (Rp13). <p>S3 is a protein of 209 to 559 amino-acid residues. As signature patterns, we selected a conserved region located in the C-terminal section.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [GSTA]-[KR]-x(6)-G-x-[LIVMT]-x(2)-[NQSCH]-x(1,3)-[LIVFCA]-x(3)-[LIV]-[DENQ]-x(7)-[LMT]-x(2)-G-x(2)-[GS]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for</p>

971

			<p>some mitochondrial S3. Other sequence(s) detected in SWISS-PROT NONE. Expert(s) to contact by email Hallick R.B. hallick@arizona.edu</p> <p>Last update December 1999 / Pattern and text revised. References [1] Otaka E., Hashimoto T., Mizuta K. Protein Seq Data Anal. 5:285-300(1993).</p>
Ribosomal_S 3_N	PDOC00474	Ribosomal protein S3 signature	<p>Ribosomal protein S3 is one of the proteins from the small ribosomal subunit. In <i>Escherichia coli</i>, S3 is known to be involved in the binding of initiator Met-tRNA. It belongs to a family of ribosomal proteins which, on the basis of sequence similarities [1], groups.</p> <ul style="list-style-type: none"> - Eubacterial S3. - Algal and plant chloroplast S3. - Cyanelle S3. - Archaeobacterial S3 - Plant mitochondrial S3. - Vertebrate S3. - Insect S3. - <i>Caenorhabditis elegans</i> S3 (C23G10.3) - Yeast S3 (Rp13). <p>S3 is a protein of 209 to 559 amino-acid residues. As signature patterns, we selected a conserved region located in the C-terminal section.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [GSTA]-[KR]-x(6)-G-x-[LIVMT]-x(2)-[NQSCH]-x(1.3)-[LIVFCA]-x(3)-[LIV]-[DENQ]-x(7)-[LMT]-x(2)-G-x(2)-[GS] Sequences known to belong to this class detected by the pattern ALL, except for some mitochondrial S3. Other sequence(s) detected in SWISS-PROT NONE Expert(s) to contact by email Hallick R.B. hallick@arizona.edu</p> <p>Last update December 1999 / Pattern and text revised References [1] Otaka E., Hashimoto T., Mizuta K. Protein Seq. Data Anal. 5:285-300(1993)</p>
RimM		RimM	<p>Accession number: PF01782 Definition: RimM Author: Bateman A Alignment method of seed: Clustalw Source of seed members: PSI-BLAST P51419 Gathering cutoffs: 25 25 Trusted cutoffs: 49.00 49.00 Noise cutoffs: -66.10 -66.10 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98083058 Reference Title: RimM and RbfA are essential for efficient processing of 16S Reference Title: rRNA in <i>Escherichia coli</i>. Reference Author: Bylund GO, Wipemo LC, Lundberg LA, Wikstrom PM; Reference Location: J Bacteriol 1998;180:73-82. Database Reference: INTERPRO; IPR002676; Comment: The RimM protein is essential for efficient processing of 16S rRNA [1]. Comment: The RimM protein was shown to have affinity for free ribosomal 30S Comment: subunits but not for 30S subunits in the 70S ribosomes [1]. Number of members: 14</p>

[illegible]

973

RNA_pol	PDOC00410	Bacteriophage-type RNA polymerase family active site signatures	<p>Many forms of RNA polymerase (EC 2.7.7.6) are known. Most RNA polymerases are multimeric enzymes, but there is a family of single chain polymerases, which are evolutionary related, and which originate from bacteriophages or from mitochondria. The RNA polymerases that belong to this family are [1]</p> <ul style="list-style-type: none"> - Podoviridae bacteriophages T3, T7, and K11 polymerase. - Bacteriophage SP6 polymerase. - Vertebrate mitochondrial polymerase (gene POLRMT). - Fungal mitochondrial polymerase (gene RPO41) - Polymerases encoded on mitochondrial linear DNA plasmids in various fungi and plants: <i>Agaricus bitorquis</i> pEM, <i>Claviceps purpurea</i> pClK1, <i>Neurospora crassa</i> Kalilo; <i>Neurospora intermedia</i> Maranh and maize S-2). <p>Two conserved aspartate and one lysine residue have been shown [2,3] to be part of the active site of T7 polymerase. We have used the regions around the first aspartate and around the lysine as signature patterns for this family of polymerases.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern P-[LIVM]-x(2)-D-[GA]-[ST]-[AC]-[SN]-[GA]-[LIVMFY]-Q [D is the active site residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [LIVMF]-x-R-x(3)-K-x(2)-[LIVMF]-M-[PT]-x(2)-Y [K is the active site residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE</p> <p>Last update July 1999 / Text revised References [1] McAllister W.T., Raskin C.A. Mol Microbiol. 10:1-6(1993).</p> <p>[2] Maksimova T.G., Mustayev A.A., Zaychikov E.F., Lyakhov D.L., Tunitskaya V.L., Akbarov A.K., Luchin S.V., Rechinsky V.O., Chernov B.K., Kochetkov S.N. Eur. J. Biochem 195 841-847(1991).</p> <p>[3] Sousa R., Chung Y.J., Rose J.P., Wang B.-C Nature 364:593-599(1993)</p>
RNA_pol_A		RNA polymerase alpha subunit	<p>Accession number: PF00623 Definition: RNA polymerase alpha subunit Author: Bateman A Alignment method of seed: HMM built from alignment Source of seed members: Pfam-B_3 (release 2.1) Gathering cutoffs: 9 0 Trusted cutoffs: 13 50 2.90 Noise cutoffs: 8.50 8.50 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97066998 Reference Title: Structural modules of the large subunits of RNA polymerase. Reference Title: Introducing archaeobacterial and chloroplast split sites in the beta and beta' subunits of Escherichia coli RNA polymerase. Reference Title: polymerase. Reference Author: Severinov K, Mustaev A, Kukarin A, Muzzin O, Bass I, Darst Reference Author: SA, Goldfarb A; Reference Location: J Biol Chem 1996;271 27969-27974. Database Reference: INTERPRO; IPR000722, Database reference: PFAM: PB003218; Comment: -!- RNA polymerases catalyse the DNA dependent polymerisation Comment: of RNA. Prokaryotes contain a single RNA polymerase</p>

974

			<p>Comment: compared to three in eukaryotes (not including mitochondrial</p> <p>Comment: and chloroplast polymerases).</p> <p>Comment: -!- Members of this family include:</p> <p>Comment: A subunit from eukaryotes</p> <p>Comment: gamma subunit from cyanobacteria</p> <p>Comment: beta' subunit from eubacteria</p> <p>Comment: A' subunit from archaeobacteria</p> <p>Comment: B' from chloroplasts</p> <p>Number of members: 202</p>
RNA_pol_A2		RNA polymerase A/beta'/A' subunit	<p>Accession number: PF01854</p> <p>Definition: RNA polymerase A/beta'/A' subunit</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_288 (release 4.2)</p> <p>Gathering cutoffs: -120 -120</p> <p>Trusted cutoffs: -116.50 -116.50</p> <p>Noise cutoffs: -125.00 -125.00</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 88335550</p> <p>Reference Title: Relatedness of archaeobacterial RNA polymerase core subunits</p> <p>Reference Title: to their eubacterial and eukaryotic equivalents</p> <p>Reference Author: Berghofer B, Krockel L, Kortner C, Truss M.</p> <p>Reference Author: Schallenberg J,</p> <p>Reference Author: Klein A;</p> <p>Reference Location: Nucleic Acids Res 1988;16:8113-8128</p> <p>Database Reference: INTERPRO, IPR002879,</p> <p>Database reference: PFAMB; PB000546;</p> <p>Database reference: PFAMB; PB000846;</p> <p>Database reference: PFAMB; PB000984;</p> <p>Database reference: PFAMB; PB001168;</p> <p>Comment: RNA polymerases catalyse the DNA dependent polymerisation</p> <p>Comment: of RNA. Prokaryotes contain a single RNA polymerase</p> <p>Comment: compared to three in eukaryotes (not including mitochondrial.</p> <p>Comment: and chloroplast polymerases).</p> <p>Comment: This family includes a region of about 400 amino acids.</p> <p>Comment: This family includes the whole archaeobacterial A' subunit, but only the C terminal region of the A subunit from eukaryotes</p> <p>Comment: and the beta' subunit from eubacteria.</p> <p>Number of members: 105</p>
RNB	PDOC00904	Ribonuclease II family signature	<p>On the basis of sequence similarities, the following bacterial and eukaryotic proteins seem to form a family.</p> <ul style="list-style-type: none"> - Escherichia coli and related bacteria ribonuclease II (EC 3.1.13.1) (RNase II) (gene rnb) [1]. RNase II is an exonuclease involved in mRNA decay. It degrades mRNA by hydrolyzing single-stranded polynucleotides processively in the 3' to 5' direction. - Bacterial ribonuclease R [2], a 3'-5' exoribonuclease that participates in an essential cell function. - Yeast protein SSD1 (or SRK1) which is implicated in the control of the cell cycle G1 phase. - Yeast protein DIS3 [3], which binds to ran (GSP1) and enhances the nucleotide-releasing activity of RCC1 on ran - Fission yeast protein dis3, which is implicated in mitotic control. - Neurospora crassa cyt-4, a mitochondrial protein required for RNA 5' and 3' end processing and splicing. - Yeast protein MSU1, which is involved in mitochondrial biogenesis. - Synechocystis strain PCC 6803 protein zam [4], which control resistance to the carbonic anhydrase inhibitor acetazolamide. - Caenorhabditis elegans hypothetical protein F48E8.6. <p>The size of these proteins range from 644 residues (rnb) to 1250 (SSD1). While their sequence is highly divergent they share a conserved domain in their C-terminal section [5]. It is possible that this domain plays a role in a putative exonuclease function that would be common to all these proteins. We have developed a signature pattern based on the core of this conserved</p>

			<p>domain</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [HI]-[FYE]-[GSTAM]-[LIVM]-x(4,5)-Y-[STALV]-x-[FWVAC]-[TV]-[SA]-P-[LIVMA]-[RQ]-[KR]-[FY]-x-D-x(3)-[HQ] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update December 1999 / Pattern and text revised. References [1] Zilhao R., Camelo L., Arraiano C.M. Mol Microbiol. 8:43-51(1993)</p> <p>[2] Cheng Z.-F., Zuo Y., Li Z., Rudd K.E., Deutscher M.P. J. Biol. Chem. 273:14077-14080(1998).</p> <p>[3] Noguchi E., Hayashi N., Azuma Y., Seki T., Nakamura M., Nakashima N., Yanagida M., He X., Mueller U., Sazer S., Nishimoto T EMBO J 15:5595-5605(1996).</p> <p>[4] Beuf L., Bedu S., Cami B., Joset F. Plant Mol. Biol. 27:779-788(1995)</p> <p>[5] Mian I.S. Nucleic Acids Res. 25 3187-3195(1997).</p>
RRF		Ribosome recycling factor	<p>Accession number PF01765 Definition: Ribosome recycling factor Author: Bashton M. Bateman A Alignment method of seed: Clustalw Source of seed members Pfam-B_949 (release 4.2) Gathering cutoffs: -35 -35 Trusted cutoffs: -34.90 -34.90 Noise cutoffs: -76.20 -76.20 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 94240115 Reference Title: Ribosome recycling factor (ribosome releasing factor) is essential for bacterial growth. Reference Author: Janosi L, Shimizu I, Kaji A; Reference Location: Proc Natl Acad Sci U S A 1994;91:4249-4253. Database Reference: INTERPRO; IPR002661. Comment: The ribosome recycling factor (RRF / ribosome release factor) dissociates Comment: the ribosome from the mRNA after termination of translation, and is Comment: essential bacterial growth [1]. Thus ribosomes are "recycled" and ready Comment: for another round of protein synthesis. Number of members: 27</p>
rve		Integrase core domain	<p>Accession number PF00665 Definition: Integrase core domain Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_10 (release 2.1) Gathering cutoffs: 9.3 9.3 Trusted cutoffs: 9.30 9.30 Noise cutoffs: 9.20 9.20 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 95099322 Reference Title: Crystal structure of the catalytic domain of HIV-1 Reference Title: integrase: similarity to other polynucleotidyl transferases</p>

			Reference Title: [see comments]
			Reference Author. Dyda F, Hickman AB, Jenkins TM, Engelman A, Craigie R.
			Reference Author. Davies DR;
			Reference Location: Science 1994,266:1981-1986.
			Database Reference: SCOP, 2ltg; fa; [SCOP-USA][CATH-PDBSUM]
			Database Reference: INTERPRO, IPR001584;
			Database Reference: PDB; 1cxu A; 56; 198;
			Database Reference: PDB, 1vsh ; 54, 199;
			Database Reference: PDB, 1vsi : 54; 199;
			Database Reference: PDB: 1vsj : 54; 199,
			Database Reference: PDB; 1cxq A; 53, 198;
			Database Reference: PDB; 1a5v ; 54: 199,
			Database Reference: PDB; 1a5w : 54; 199;
			Database Reference: PDB; 1a5x ; 54, 199,
			Database Reference: PDB; 1asv ; 54, 199;
			Database Reference: PDB; 1vsm A; 54; 199;
			Database Reference: PDB, 1czb A; 53; 198:
			Database Reference: PDB; 1asw . 53; 201;
			Database Reference: PDB, 1cz9 A. 59; 197;
			Database Reference: PDB; 1vsk . 54; 199;
			Database Reference: PDB; 1vsl A; 54; 199,
			Database Reference: PDB 1asu ; 53; 207,
			Database Reference: PDB; 1c0m A; 53; 213;
			Database Reference: PDB: 1vsd : 54, 88;
			Database Reference: PDB; 1vse : 54; 88;
			Database Reference: PDB; 1c1a B: 55, 213,
			Database Reference: PDB; 1c0m B; 54, 213;
			Database Reference: PDB; 1c0m D: 54, 213;
			Database Reference: PDB; 1c1a A; 53, 213;
			Database Reference: PDB; 1c0m C: 53; 213;
			Database Reference: PDB; 1bhl ; 57; 201;
			Database Reference: PDB; 1bi4 B; 57, 201;
			Database Reference: PDB; 1bi3 B; 57, 201;
			Database Reference: PDB; 1b9f A; 56, 201;
			Database Reference: PDB; 1bis B; 56, 201;
			Database Reference: PDB; 1qs4 B; 56, 201;
			Database Reference: PDB; 1qs4 C; 56, 201;
			Database Reference: PDB; 1biz A; 54; 201;
			Database Reference: PDB; 1itg ; 55; 201;
			Database Reference: PDB, 1bi4 C, 53; 201;
			Database Reference: PDB, 1bi3 C, 53; 201;
			Database Reference: PDB, 2itg : 53; 201;
			Database Reference: PDB; 1b9d A; 57; 189,
			Database Reference: PDB; 1bi4 A, 57; 201,
			Database Reference: PDB, 1bi3 A, 57; 201;
			Database Reference: PDB; 1bis A; 56; 201,
			Database Reference: PDB; 1biu A; 56, 201;
			Database Reference: PDB; 1biu B; 56; 201;
			Database Reference: PDB, 1biu C; 56; 201;
			Database Reference: PDB, 1qs4 A; 56; 201;
			Database Reference: PDB; 1b92 A; 56, 201;
			Database Reference: PDB; 1biz B; 58; 201,
			Database Reference: PDB; 1b9d A; 382; 390;
			Database Reference: PDB; 1wjb A; 53; 55;
			Database Reference: PDB; 1wjb B; 53; 55;
			Database Reference: PDB, 1wjd A; 53; 55,
			Database Reference: PDB; 1wjd B; 53; 55;
			Database Reference: PDB, 1wjf A; 53; 55;
			Database Reference: PDB: 1wjf B; 53, 55;
			Database reference: PFAMB; PB000048;
			Database reference: PFAMB; PB007709;
			Database reference: PFAMB; PB013923;
			Database reference: PFAMB; PB013938;
			Database reference: PFAMB, PB018509;
			Database reference: PFAMB; PB020302;
			Database reference: PFAMB; PB025327;
			Database reference: PFAMB; PB028352;
			Database reference: PFAMB; PB032740;
			Database reference: PFAMB; PB040612;
			Database reference: PFAMB; PB040636;
			Database reference: PFAMB, PB040684;
			Database reference: PFAMB; PB040695;
			Database reference: PFAMB; PB040730;

977

			<p>Database reference: PFAMB; PB040824; Database reference: PFAMB; PB041112; Database reference: PFAMB; PB041143; Database reference: PFAMB; PB041275; Database reference: PFAMB; PB041356; Database reference: PFAMB; PB041375; Database reference: PFAMB; PB041456; Database reference: PFAMB; PB041459; Database reference: PFAMB; PB041522; Database reference: PFAMB; PB041665; Database reference: PFAMB; PB041761; Database reference: PFAMB; PB041816; Database reference: PFAMB; PB041885; Comment: Integrase mediates integration of a DNA copy of the viral genome into the host chromosome. Integrase is composed of Comment: three domains. The amino-terminal domain is a zinc binding Comment: domain Integrase_Zn. This domain is the central catalytic domain. The carboxyl terminal domain that is a non-specific Comment: DNA binding domain integrase. Comment: The catalytic domain acts as an endonuclease when two nucleotides are removed from the 3' ends of the blunt-ended Comment: viral DNA made by reverse transcription. This domain also catalyses the DNA strand transfer reaction of the 3' ends of the viral DNA to the 5' ends of the integration site [1]. Comment: Number of members: 1147</p>
S4		S4 domain	<p>Accession number. PF01479 Definition: S4 domain Author: Bateman A Alignment method of seed: Clustalw Source of seed members. Medline:99193178 Gathering cutoffs: 17 17 Trusted cutoffs 17.20 17 20 Noise cutoffs: 16.70 16.70 HMM build command line hmmbuild HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 99193178 Reference Title: Novel predicted RNA-binding domains associated with the Reference Title: translation machinery Reference Author: Aravind L, Koonin EV; Reference Location: J Mol Evol 1999;48:291-302. Reference Number: [2] Reference Medline: 98372721 Reference Title: The crystal structure of ribosomal protein S4 reveals a two-domain molecule with an extensive RNA-binding surface: Reference Title: one domain shows structural homology to the ETS DNA-binding Reference Title: motif. Reference Author: Davies C, Gerstner RB. Draper DE, Ramakrishnan V. White SW; Reference Location: EMBO J 1998;17 4545-4558. Database Reference: SCOP; 1c06; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR002942; Database Reference: PDB; 1c05 A; 51; 98; Database Reference: PDB; 1c06 A; 51; 98; Database Reference: PDB; 1dm9 A; 9; 55; Database Reference: PDB; 1dm9 B; 9; 55; Database reference: PFAMB; PB001751; Database reference: PFAMB; PB041147; Database reference: PFAMB; PB041148; Comment: The S4 domain is a small domain consisting of 60-65 amino acid residues Comment: that was detected in the bacterial ribosomal protein S4, eukaryotic Comment: ribosomal S9, two families of pseudouridine synthases. a novel family Comment: of predicted RNA methylases, a yeast protein containing a</p>

978

			<p>pseudouridine Comment: synthetase and a deaminase domain, bacterial tyrosyl-tRNA synthetases, Comment: and a number of uncharacterized, small proteins that may be involved in Comment: translation regulation [1]. The S4 domain probably mediates binding to Comment: RNA. Number of members: 256</p>
SAA proteins	PDOC00762	Serum amyloid A proteins signature	<p>The serum amyloid A (SAA) proteins comprise a family of vertebrate proteins that associate predominantly with high density lipoproteins (HDL) [1,2]. The synthesis of certain members of the family is greatly increased (as much as a 1000 fold) in inflammation, thus making SAA a major acute phase reactant. While the major physiological function of SAA is unclear, prolonged elevation of plasma SAA levels, as in chronic inflammation, however, results in a pathological condition, called amyloidosis, which affects the liver, kidney and spleen and which is characterized by the highly insoluble accumulation of SAA in these tissues.</p> <p>SAA are proteins of about 110 amino acid residues. As a signature pattern, we selected the most highly conserved region, which is located in the central part of the sequence.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern A-R-G-N-Y-[ED]-A-x-[QKR]-R-G-x-G-G-x-W-A Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE. Last update June 1994 / First entry. References [1] Malle E., Steinmetz A., Raynes J G. Atherosclerosis 102:131-146(1993) [2] Uhlir C.M., Burgess C J., Sharp P.M., Whitehead A.S. Genomics 19:228-235(1994).</p>
SAM		SAM domain (Sterile alpha motif)	<p>Accession number: PF00536 Definition: SAM domain (Sterile alpha motif) Author: Bateman A Alignment method of seed: Clustalw Source of seed members [1],[2] Gathering cutoffs: 11 0 Trusted cutoffs: 11 00 3.70 Noise cutoffs: 10 90 10.90 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 96100659 Reference Title: SAM: A novel motif in yeast sterile alpha and Drosophila polyhomeotic proteins Reference Author: Ponting CP; Reference Location: Prot Sci 1995 4:1928-1930. Reference Number: [2] Reference Medline: 97160498 Reference Title: SAM as a protein interaction domain involved in developmental regulation. Reference Author: Shultz J, Ponting CP, Hofmann K, Bork P; Reference Location: Prot Sci 1997;6:249-253. Reference Number: [3] Reference Medline: 99101382 Reference Title: The crystal structure of an Eph receptor SAM domain reveals Reference Title: a mechanism for modular dimerization. Reference Author: Stapleton D, Balan I, Pawson T, Sicheri F, Reference Location: Nat Struct Biol 1999;6:44-49. Database reference: SMART; SAM; Database Reference: SCOP, 1b0x; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO, IPR001660;</p>

979

			<p>Database Reference PDB; 1b0x A; 910; 973;</p> <p>Database Reference PDB; 1sgg ; 7; 70;</p> <p>Database Reference PDB; 1b4f A; 7; 71;</p> <p>Database Reference PDB; 1b4f C; 7; 71;</p> <p>Database Reference PDB; 1b4f E; 7; 71;</p> <p>Database Reference PDB; 1b4f D; 7; 71;</p> <p>Database Reference PDB; 1b4f H; 7; 71;</p> <p>Database Reference PDB; 1b4f F; 7; 71;</p> <p>Database Reference PDB; 1b4f G; 7; 71;</p> <p>Database Reference PDB; 1b4f B; 7; 71;</p> <p>Database reference: PFAMB; PB008631;</p> <p>Database reference: PFAMB; PB040678;</p> <p>Database reference: PFAMB; PB041111;</p> <p>Database reference: PFAMB; PB041385;</p> <p>Comment: It has been suggested that SAM is an evolutionarily conserved protein</p> <p>Comment: binding domain that is involved in the regulation of numerous</p> <p>Comment: developmental processes in diverse eukaryotes.</p> <p>Comment: The SAM domain can potentially function as a protein interaction</p> <p>Comment: module through its ability to homo- and heterooligomerise with</p> <p>Comment: other SAM domains</p> <p>Number of members: 110</p>
SAM decarbox		Adenosylmethionine decarboxylase	<p>Accession number: PF01536</p> <p>Definition: Adenosylmethionine decarboxylase</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_600 (release 4.0)</p> <p>Gathering cutoffs: 11 11</p> <p>Trusted cutoffs 17 90 17 90</p> <p>Noise cutoffs: 5.70 5.70</p> <p>HMM build command line: hmmbuild -f HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 98098079</p> <p>Reference Title: Cloning, mapping and mutational analysis of the</p> <p>Reference Title: S-adenosylmethionine decarboxylase gene in <i>Drosophila melanogaster</i>.</p> <p>Reference Author: Larsson J, Rasmuson-Lestander A,</p> <p>Reference Location: Mol Gen Genet 1997;256:652-660</p> <p>Database Reference: SCOP; 1jen; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR001985;</p> <p>Database Reference: PDB; 1jen C; 69; 328;</p> <p>Database Reference: PDB; 1jen A; 69; 329;</p> <p>Database Reference: PDB; 1jen B; 4; 67;</p> <p>Database Reference: PDB; 1jen D; 5; 66;</p> <p>Comment: This is a family of S-adenosylmethionine decarboxylase (SAMDC) proenzymes.</p> <p>Comment: In the biosynthesis of polyamines SAMDC produces decarboxylated</p> <p>Comment: S-adenosylmethionine, which serves as the aminopropyl moiety necessary</p> <p>Comment: for spermidine and spermine biosynthesis from putrescine</p> <p>Comment: [1] The Pfam</p> <p>Comment: alignment contains both the alpha and beta chains that are cleaved to</p> <p>Comment: form the active enzyme.</p> <p>Number of members: 34</p>
SBF		Sodium Bile acid symporter family	<p>Accession number: PF01758</p> <p>Definition: Sodium Bile acid symporter family</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_697 (release 4.2)</p> <p>Gathering cutoffs: -19 -19</p> <p>Trusted cutoffs: -12.50 -12.50</p> <p>Noise cutoffs: -26.40 -26.40</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97377989</p>

980

981

			<p>Reference Title: 1H NMR assignment and global fold of napin Bnlb, a representative 2S albumin seed protein.</p> <p>Reference Author: Rico M, Bruix M, Gonzalez C, Monsalve RI, Rodriguez R;</p> <p>Reference Location: Biochemistry 1996;35:15672-15682.</p> <p>Database Reference: SCOP, 1pnb; fa. [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR000617;</p> <p>Database reference: PFAMB; PB029622;</p> <p>Comment: Members of this family are composed of two chains (both included in the alignment), these are co-translated and later cleaved.</p> <p>Comment: The two chains are disulphide linked together.</p> <p>Number of members: 27</p>
SH2	PDOC50001	Src homology 2 (SH2) domain profile	<p>The Src homology 2 (SH2) domain is a protein domain of about 100 amino-acid residues first identified as a conserved sequence region between the oncoproteins Src and Fps [1]. Similar sequences were later found in many other intracellular signal-transducing proteins [2]. SH2 domains function as regulatory modules of intracellular signalling cascades by interacting with high affinity to phosphotyrosine-containing target peptides in a sequence-specific and strictly phosphorylation-dependent manner [3,4,5,6].</p> <p>The SH2 domain has a conserved 3D structure consisting of two alpha helices and six to seven beta-strands. The core of the domain is formed by a continuous beta-meander composed of two connected beta-sheets [7].</p> <p>So far, SH2 domains have been identified in the following proteins:</p> <ul style="list-style-type: none"> - Many vertebrate, invertebrate and retroviral cytoplasmic (non-receptor) protein tyrosine kinases. In particular in the Src, Abl, Btk, Csk and ZAP70 families of kinases. - Mammalian phosphatidylinositol-specific phospholipase C gamma-1 and -2. <p>Two copies of the SH2 domain are found in those proteins in between the catalytic 'X-' and 'Y-boxes' (see <PDOC50007>)</p> <ul style="list-style-type: none"> - Mammalian phosphatidylinositol 3-kinase regulatory p85 subunit. - Some vertebrate and invertebrate protein-tyrosine phosphatases. - Mammalian Ras GTPase-activating protein (GAP). - Adaptor proteins mediating binding of guanine nucleotide exchange factors to growth factor receptors: vertebrate GRB2, Caenorhabditis elegans sem-5 and Drosophila DRK. - Mammalian Vav oncoprotein, a guanine-nucleotide exchange factor of the CDC24 family. - Miscellaneous proteins interacting with vertebrate receptor protein tyrosine kinases: oncoprotein Crk, mammalian cytoplasmic proteins Nck, Shc. - STAT proteins (signal transducers and activators of transcription). - Chicken tensin. - Yeast transcriptional control protein SPT6. <p>The profile developed to detect SH2 domains is based on a structural alignment consisting of 8 gap-free blocks and 7 linker regions totaling 92 match positions.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Sequences known to belong to this class detected by the profile ALL</p> <p>Other sequence(s) detected in SWISS-PROT protein tyrosine kinases JAK1 and JAK2.</p> <p>Expert(s) to contact by email Zvelebil M. marketa@ludwig.ucl.ac.uk</p> <p>Last update November 1995 / First entry.</p> <p>References [1] Sadowski I., Stone J.C., Pawson T. Mol. Cell. Biol. 6:4396-4408(1986).</p> <p>[2] Russell R.B., Breed J., Barton G.J.</p>

			<p>FEBS Lett. 304:15-20(1992).</p> <p>[3] Marangere L E M., Pawson T. J. Cell Sci. Suppl. 18:97-104(1994).</p> <p>[4] Pawson T., Schlessinger J Curr Biol. 3:434-442(1993).</p> <p>[5] Mayer B.J. Baltimore D. Trends Cell. Biol. 3:8-13(1993).</p> <p>[6] Pawson T. Nature 373:573-580(1995)</p> <p>[7] Kuriyan J., Cowburn D Curr. Opin. Struct Biol 3:828-837(1993)</p>
Shikimate_D H		Shikimate / quininate 5- dehydrogenase	<p>Accession number: PF01488 Definition: Shikimate / quininate 5-dehydrogenase Author: Bashton M. Bateman A Alignment method of seed Clustalw Source of seed members Pfam-B_336 (release 4.0) Gathering cutoffs: -50 -50 Trusted cutoffs -48.00 -48.00 Noise cutoffs: -82.00 -82.00 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 96048023 Reference Title: The molecular biology of multidomain proteins. Selected Reference Title: examples. Reference Author: Hawkins AR, Lamb HK; Reference Location: Eur J Biochem 1995;232:7-18. Database Reference: INTERPRO, IPR002907; Comment: This family contains both shikimate and quininate dehydrogenases. Comment: Shikimate 5-dehydrogenase catalyses the conversion of Comment: shikimate to 5-dehydroshikimate. This reaction is part of Comment: the shikimate pathway which is involved in the biosynthesis Comment: of aromatic amino acids. Comment: Quinate 5-dehydrogenase catalyses the conversion of Comment: quinate to 5-dehydroquinate This reaction is part of Comment: the quinate pathway where quinic acid is exploited as Comment: a source of carbon in prokaryotes and microbial Comment: eukaryotes. Comment: Both the shikimate and quinate pathways share two common Comment: pathway metabolites 3-dehydroquinate and dehydroshikimate. Number of members: 58</p>
Sigma54_fact ors	PDOC00593	Sigma-54 factors family signatures and profile	<p>Sigma factors [1] are bacterial transcription initiation factors that promote the attachment of the core RNA polymerase to specific initiation sites and are then released. They alter the specificity of promoter recognition. Most bacteria express a multiplicity of sigma factors. Two of these factors, sigma-70 (gene rpoD), generally known as the major or primary sigma factor, and sigma-54 (gene rpoN or ntrA) direct the transcription of a wide variety of genes. The other sigma factors, known as alternative sigma factors, are required for the transcription of specific subsets of genes.</p> <p>With regard to sequence similarity, sigma factors can be grouped into two classes: the sigma-54 and sigma-70 families. The sigma-70 family has many different sigma factors (see the relevant entry <PDOC00592>). The sigma-54 family consists exclusively of sigma-54 factor [2,3] required for the transcription of promoters that have a characteristic -24 and -12 consensus recognition element but which are devoid of the typical -10,-35 sequences recognized by the major sigma factors. The sigma-54 factor is also characterized by its interaction with ATP-dependent positive regulatory</p>

			<p>proteins that bind to upstream activating sequences</p> <p>Structurally sigma-54 factors consist of three distinct regions</p> <ul style="list-style-type: none"> - A relatively well conserved N-terminal glutamine-rich region of about 50 residues that contains a potential leucine zipper motif. - A region of variable length which is not well conserved - A well conserved C-terminal region of about 350 residues that contains a second potential leucine zipper, a potential DNA-binding 'helix-turn-helix' motif and a perfectly conserved octapeptide whose function is not known <p>We developed two signature patterns for this family of sigma factors. The first starts two residues before the N-terminal extremity of the helix-turn-helix region and ends two residues before its C-terminal extremity. The second is the conserved octapeptide. A profile has also been designed that covers the whole C-terminal region.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern P-[LIVM]-x-[LIVM]-x(2)-[LIVM]-A-x(2)-[LIVMFT]-x(2)-[HS]-x-S-T-[LIVM]-S-R Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE</p> <p>Consensus pattern R-R-T-[IV]-[ATN]-K-Y-R Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Sequences known to belong to this class detected by the profile ALL. Other sequence(s) detected in SWISS-PROT NONE</p> <p>Note this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so</p> <p>Last update July 1999 / Patterns and text revised.</p> <p>References [1] Hermann J D., Chamberlin M.J. Annu. Rev. Biochem. 57:839-872(1988).</p> <p>[2] Thoeny B., Hennecke H. FEMS Microbiol. Rev 5 341-358(1989).</p> <p>[3] Merrick M.J. Mol. Microbiol. 10 903-909(1993).</p>
SLH	PDOC00823	S-layer homology domain signature	<p>S-layers are paracrystalline mono-layered assemblies of (glyco)proteins which coat the surface of bacteria [1]. Several S-layer proteins and some other cell wall proteins contain one or more copies of a domain of about 50-60 residues, which has been called SLH (for S-layer homology) [2]. There is strong evidence that this domain serves as an anchor to the peptidoglycan [3]. The SLH domain has been found in:</p> <ul style="list-style-type: none"> - S-layer glycoprotein of <i>Acetogenium kivui</i> (3 copies). - S-layer 125 Kd protein of <i>Bacillus sphaericus</i> (3 copies). - S-layer protein of <i>Bacillus anthracis</i> (3 copies). - S-layer protein of <i>Bacillus licheniformis</i> (3 copies). - S-layer protein (HWP) from <i>Bacillus brevis</i> strain HPD31 (3 copies). - Middle cell wall protein (MWP) from <i>Bacillus brevis</i> strain 47 (3 copies). - S-layer protein (p100) of <i>Thermus thermophilus</i> (1 copy). - Outer membrane protein Omp-alpha from <i>Thermotoga maritima</i> (1 copy). - Cellulosome anchoring protein (gene <i>ancA</i>), outer layer protein B (OlpB) and a further potential cell surface glycoprotein from <i>Clostridium thermocellum</i> (3 copies; the first copy is missing its N-terminal third which is appended to the end of the third copy; may have arisen by circular permutation). - Amylopullulanase (gene <i>amyB</i>) from <i>Thermoanaerobacter thermosulfurogenes</i> (3 copies) - Amylopullulanase (gene <i>aapT</i>) from <i>Bacillus</i> strain XAL-601 (3 copies)

		<p>- Endoglucanase from <i>Bacillus</i> strain KSM-635 (3 copies). - Exoglucanase (gene <i>xynX</i>) from <i>Clostridium thermocellum</i> (3 copies). - Xylanase A (gene <i>xynA</i>) from <i>Thermoanaerobacter saccharolyticum</i> (2 copies; 3 copies if a frameshift is taken into account). - Protein involved in butirosin production (ButB) from <i>Bacillus circulans</i> (2 incomplete copies; 3 copies if three frameshifts are taken into account). - Two hypothetical proteins from <i>Synechocystis</i> strain PCC 6803 (1 copy each). - A hypothetical protein with sequence similarity to amylopullulanases found 3' of amylase gene from <i>Bacillus circulans</i> (fragment of 1 copy; 3 copies if two frameshifts are taken into account).</p> <p>SLH domains are found at the N- or C-termini of mature proteins. They occur in single copy followed by a predicted coiled coil domain, or in three contiguous copies. Structurally, the SLH domain is predicted to contain two alpha-helices flanking a beta strand. The SLH sequences are fairly divergent with an average identity of about 25%. It is however possible to build a sequence pattern that starts at the second position of the domain and that spans 3/4 of its length.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LVFYT]-x-[DA]-x(2,5)-[DNGSATPHY]-[FYWPDA]-x(4)-[LIV]-x(2)-[GTALV]-x(4,6)-[LIVFYC]-x(2)-G-x-[PGSTA]-x(2,3)-[MFYA]-x-[PGAV]-x(3,10)-[LIVMA]-[STKR]-[RY]-x-[EQ]-x-[STALIVM] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE Expert(s) to contact by email Lupas A.N. lupas@vms.biochem.mpg.de</p> <p>Last update November 1997 / Pattern and text revised.</p> <p>References [1] Beveridge T.J Curr. Opin. Struct. Biol. 4:204-212(1994).</p> <p>[2] Lupas A., Engelhardt H., Peters J., Santarius U., Volker S., Baumeister W. J. Bacteriol. 176:1224-1233(1994).</p> <p>[3] Lemaire M., Ohayon H., Gounon P., Fujino T., Beguin P. J. Bacteriol. 177:2451-2459(1995).</p>
Smr		<p>Smr domain</p> <p>Accession number PF01713 Definition: Smr domain Author: Bateman A Alignment method of seed: Clustalw Source of seed members: [1] Gathering cutoffs: 0 0 Trusted cutoffs 1.40 1.40 Noise cutoffs -7.90 -7.90 HMM build command line: <code>hmmbuild HMM SEED</code> HMM build command line: <code>hmmcalibrate --seed 0 HMM</code> Reference Number: [1] Reference Medline: 10431172 Reference Title: Smr. a bacterial and eukaryotic homologue of the C-terminal region of the MutS2 family. Reference Author: Moreira D, Philippe H; Reference Location: Trends Biochem Sci 1999;24:298-300. Database Reference: INTERPRO; IPR002625. Comment: This family includes the Smr (Small MutS Related) proteins, Comment: and the C-terminal region of the MutS2 protein. It has been Comment: suggested that this domain interacts with the MutS1 Comment: Swiss:P23909 protein in the case of Smr proteins and with Comment: the N-terminal MutS related region of MutS2 Swiss:P94545 [1]. Number of members: 14</p>

SRF-TF	PDO00302	MADS-box domain signature and profile	<p>A number of transcription factors contain a conserved domain of 56 amino-acid residues, sometimes known as the MADS-box domain [E1]. They are listed below:</p> <ul style="list-style-type: none"> - Serum response factor (SRF) [1], a mammalian transcription factor that binds to the Serum Response Element (SRE). This is a short sequence of dyad symmetry located 300 bp to the 5' end of the transcription initiation site of genes such as c-fos. - Mammalian myocyte-specific enhancer factors 2A to 2D (MEF2A to MEF2D). These proteins are transcription factor which binds specifically to the MEF2 element present in the regulatory regions of many muscle-specific genes. - Drosophila myocyte-specific enhancer factor 2 (MEF2). - Yeast GRM/PRTF protein (gene MCM1) [2], a transcriptional regulator of mating-type-specific genes - Yeast arginine metabolism regulation protein I (gene ARGR1 or ARG80) - Yeast transcription factor RLM1. - Yeast transcription factor SMP1 - Arabidopsis thaliana agamous protein (AG) [3], a probable transcription factor involved in regulating genes that determines stamen and carpel development in wild-type flowers. Mutations in the AG gene result in the replacement of the stamens by petals and the carpels by a new flower. - Arabidopsis thaliana homeotic proteins Apetala1 (AP1), Apetala3 (AP3) and Pistillata (PI) which act locally to specify the identity of the floral meristem and to determine sepal and petal development [4]. - Antirrhinum majus and tobacco homeotic protein deficiens (DEFA) and globosa (GLO) [5]. Both proteins are transcription factors involved in the genetic control of flower development. Mutations in DEFA or GLO cause the transformation of petals into sepals and of stamens into carpels - Arabidopsis thaliana putative transcription factors AGL1 to AGL6 [6] - Antirrhinum majus morphogenetic protein DEF H33 (squamosa). <p>In SRF, the conserved domain has been shown [1] to be involved in DNA-binding and dimerization. We have derived a pattern that spans the complete length of the domain. The profile also spans the length of the MADS-box.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern R-x-[RK]-x(5)-I-x-[DNGSK]-x(3)-[KR]-x(2)-T-[FY]-x-[RK](3)-x(2)-[LIVM]-x-K(2)-A-x-E-[LIVM]-[STA]-x-L-x(4)-[LIVM]-x-[LIVM](3)-x(6)-[LIVMF]-x(2)-[FY]</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Sequences known to belong to this class detected by the profile ALL.</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note this documentation entry is linked to both signature patterns and a profile. As the profile is much more sensitive than the patterns, you should use it if you have access to the necessary software tools to do so</p> <p>Last update July 1999 / Pattern and text revised.</p> <p>References</p> <p>[1] Norman C., Runswick M., Pollock R., Treisman R. Cell 55:989-1003(1988).</p> <p>[2] Passmore S., Maine G.T., Elble R., Christ C., Tye B.-K. J. Mol. Biol. 204:593-606(1988).</p> <p>[3] Yanofsky M., Ma H., Bowman J., Drews G., Feldmann K.A., Meyerowitz E.M. Nature 346:35-39(1990).</p> <p>[4] Goto K., Meyerowitz E.M. Genes Dev. 8:1548-1560(1994).</p>
--------	----------	---------------------------------------	---

			<p>[5] Troebner W , Ramirez L., Motte P., Hue I., Huijser P., Loennig W -E., Saedler H., Sommer H., Schwartz-Sommer Z. EMBO J. 11.4693-4704(1992).</p> <p>[6] Ma H , Yanofsky M.F., Meyerowitz E.M. Genes Dev. 5:484-495(1991)</p> <p>[E1] http://transfac.gbf-braunschweig.de/cgi-bin/qt/getEntry.pl?C0014</p>
SRP19		SRP19 protein	<p>Accession number: PF01922 Definition: SRP19 protein Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: 25 25 Trusted cutoffs: 31.20 31.20 Noise cutoffs: -28 50 -28.50 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 89041541 Reference Title: Isolation and characterization of a cDNA clone encoding the Reference Title: 19 kDa protein of signal recognition particle (SRP): Reference Title: expression and binding to 7SL RNA. Reference Author: Lingelbach K, Zwieb C, Webb JR, Marshallsay C, Hoben PJ. Reference Author: Walter P, Dobberstein B, Reference Location: Nucleic Acids Res 1988;16:9431-9442 Reference Number: [2] Reference Medline: 92220168 Reference Title: SEC65 gene product is a subunit of the yeast signal recognition particle required for its integrity Reference Author: Hann BC, Stirling CJ, Walter P; Reference Location: Nature 1992;356:532-533 Reference Number: [3] Reference Medline: 92220169 Reference Title: The S. cerevisiae SEC65 gene encodes a component of yeast Reference Title: signal recognition particle with homology to human SRP19. Reference Author: Stirling CJ, Hewitt EW; Reference Location: Nature 1992;356:534-537. Database Reference: INTERPRO, IPR002778; Comment: The signal recognition particle (SRP) binds to the signal peptide of Comment: proteins as they are being translated. The binding of the SRP halts Comment: translation and the complex is then transported to the endoplasmic Comment: reticulum's cytoplasmic surface. The SRP then aids translocation of Comment: the protein through the ER membrane. The SRP is a ribonucleoprotein Comment: that is composed of a small RNA and several proteins. One of these Comment: proteins is the SRP19 protein [1] (Sec65 in yeast [2,3]). Number of members: 13</p>
SSB	PDOC00602	Single-strand binding protein family signatures	<p>The Escherichia coli single-strand binding protein [1] (gene ssb), also known as the helix-destabilizing protein, is a protein of 177 amino acids. It binds tightly, as a homotetramer, to single-stranded DNA (ss-DNA) and plays an important role in DNA replication, recombination and repair.</p> <p>Closely related variants of SSB are encoded in the genome of a variety of large self-transmissible plasmids. SSB has also been characterized in bacteria such as <i>Proteus mirabilis</i> or <i>Serratia marcescens</i>.</p> <p>Eukaryotic mitochondrial proteins that bind ss-DNA and are probably involved in mitochondrial DNA replication are structurally and evolutionary related to</p>

987

			<p>prokaryotic SSB. Proteins currently known to belong to this subfamily are listed below [2].</p> <ul style="list-style-type: none"> - Mammalian protein Mt-SSB (P16). - Xenopus Mt-SSBs and Mt-SSBr. - Drosophila MtSSB. - Yeast protein RIM1. <p>We have developed two signature patterns for these proteins. The first is a conserved region in the N-terminal section of the SSB's. The second is a centrally located region which, in <i>Escherichia coli</i> SSB, is known to be involved in the binding of DNA.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVMF]-[NST]-[KRHST]-[LIVM]-x-[LIVMF](2)-G-[NHRK]-[LIVMA]-[GST]-x-[DENT]</p> <p>Sequences known to belong to this class detected by the pattern ALL</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern T-x-W-[HY]-[RNS]-[LIVM]-x-[LIVMF]-[FY]-[NGKR]</p> <p>Sequences known to belong to this class detected by the pattern A majority</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update December 1999 / Patterns and text revised.</p> <p>References [1] Meyer R.R., Laine P S. Microbiol. Rev. 54:342-380(1990).</p> <p>[2] Stroumbakis N.D., Li Z, Tolia P.P. Gene 143.171-177(1994).</p>
START		START domain	<p>Accession number: PF01852</p> <p>Definition: START domain</p> <p>Author: SMART</p> <p>Alignment method of seed: Manual</p> <p>Source of seed members: Alignment kindly provided by SMART</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 106.20 106 20</p> <p>Noise cutoffs: -20.90 -20.90</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 99257451</p> <p>Reference Title: START: a lipid-binding domain in StAR, HD-ZIP and signalling proteins.</p> <p>Reference Author: Ponting CP, Aravind L,</p> <p>Reference Location: Trends Biochem Sci 1999;24:130-132.</p> <p>Database reference: SMART: START;</p> <p>Database Reference: INTERPRO: IPR002913,</p> <p>Number of members: 41</p>
Sterol_desat		Sterol desaturase	<p>Accession number: PF01598</p> <p>Definition: Sterol desaturase</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_905 (release 4 1)</p> <p>Gathering cutoffs: -13 -13</p> <p>Trusted cutoffs: 12.90 12.90</p> <p>Noise cutoffs: -44.50 -44.50</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 91323727</p> <p>Reference Title: Cloning, disruption and sequence of the gene encoding yeast</p> <p>Reference Title: C-5 sterol desaturase.</p> <p>Reference Author: Arthington BA, Bennett LG, Skatrud PL, Guynn CJ, Barbuch</p> <p>Reference Author: RJ, Ulbright CE, Bard M;</p>

988

			<p>Reference Location: Gene 1991,102:39-44.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 96133902</p> <p>Reference Title: Cloning and characterization of ERG25, the <i>Saccharomyces cerevisiae</i> gene encoding C-4 sterol methyl oxidase.</p> <p>Reference Author: Bard M, Bruner DA, Pierson CA, Lees ND, Biermann B, Frye L.</p> <p>Reference Author: Koegel C, Barbuch R;</p> <p>Reference Location: Proc Natl Acad Sci U S A 1996;93:186-190.</p> <p>Reference Number: [3]</p> <p>Reference Medline: 96351930</p> <p>Reference Title: Molecular characterization of the CER1 gene of <i>arabidopsis</i></p> <p>Reference Title: involved in epicuticular wax biosynthesis and pollen fertility.</p> <p>Reference Author: Aarts MG, Keijzer CJ, Stiekema WJ, Pereira A;</p> <p>Reference Location: Plant Cell 1995;7 2115-2127.</p> <p>Database Reference: INTERPRO; IPR001541;</p> <p>Database reference: PFAMB; PB041851;</p> <p>Comment: This family includes C-5 sterol desaturase and C-4 sterol methyl oxidase. Members of this family are involved in cholesterol biosynthesis and biosynthesis a plant cuticular wax These enzymes contain many conserved histidine residues Members of this family are integral membrane proteins.</p> <p>Number of members: 34</p>
Sulfate_trans p	PDOC00870	Sulfate transporters signature	<p>A number of proteins involved in the transport of sulfate across a membrane as well as some yet uncharacterized proteins have been shown [1,2] to be evolutionary related. These proteins are:</p> <ul style="list-style-type: none"> - <i>Neurospora crassa</i> sulfate permease II (gene <i>cys-14</i>). - Yeast sulfate permeases (genes <i>SUL1</i> and <i>SUL2</i>). - Rat sulfate anion transporter 1 (<i>SAT-1</i>). - Mammalian DTDST, a probable sulfate transporter which, in Human, is involved in the genetic disease, diastrophic dysplasia (DTD). - Sulfate transporters 1, 2 and 3 from the legume <i>Stylosanthes hamata</i>. - Human pendrin (gene <i>PDS</i>), which is involved in a number of hearing loss genetic diseases - Human protein DRA (Down-Regulated in Adenoma). - Soybean early nodulin 70. - <i>Escherichia coli</i> hypothetical protein <i>ychM</i>. - <i>Caenorhabditis elegans</i> hypothetical protein F41D9.5. <p>As expected by their transport function, these proteins are highly hydrophobic and seem to contain about 12 transmembrane domains. The best conserved region seems to be located in the second transmembrane region and is used as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [PAV]-x-Y-[GS]-L-Y-[STAG](2)-x(4)-[LIVFYA]-[LIVST]-[YI]-x(3)-[GA]-[GST]-S-[KR]</p> <p>Sequences known to belong to this class detected by the pattern ALL</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update July 1999 / Pattern and text revised.</p> <p>References [1] Sandal N.N., Marcker K.A. Trends Biochem. Sci. 19:19-19(1994).</p> <p>[2] Smith F.W., Hawkesford M.J., Prosser I.M., Clarkson D.T. Mol. Gen. Genet. 247:709-715(1995).</p>

989

TGT

	tRNA-ribosyltransferase	<p>Definition: Queuine tRNA-ribosyltransferase</p> <p>Author: Bashton M, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1643 (release 4.1)</p> <p>Gathering cutoffs: -132 -132</p> <p>Trusted cutoffs: -110.00 -110.00</p> <p>Noise cutoffs: -155.40 -155.40</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96256303</p> <p>Reference Title: Crystal structure of tRNA-guanine transglycosylase: RNA modification by base exchange.</p> <p>Reference Author: Romier C, Reuter K, Suck D, Ficner R;</p> <p>Reference Location: EMBO J 1996;15:2850-2857.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 93287116</p> <p>Reference Title: tRNA-guanine transglycosylase from Escherichia coli. Overexpression, purification and quaternary structure.</p> <p>Reference Author: Garcia GA, Koch KA, Chong S;</p> <p>Reference Location: J Mol Biol 1993;231:489-497.</p> <p>Database Reference: SCOP: 1pud: fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR002616,</p> <p>Database Reference: PDB; 1efz A; 138; 379;</p> <p>Database Reference: PDB; 1enu A; 138; 379;</p> <p>Database Reference: PDB; 1pud , 138; 379;</p> <p>Database Reference: PDB, 1wkd , 138; 379,</p> <p>Database Reference: PDB; 1wke ; 138, 379;</p> <p>Database Reference: PDB; 1wkf ; 138, 379;</p> <p>Database reference: PFAM; PB037884;</p> <p>Comment: This is a family of queuine tRNA-ribosyltransferases</p> <p>Comment: EC:2.4.2.29, also known as tRNA-guanine transglycosylase</p> <p>Comment: and guanine insertion enzyme.</p> <p>Comment: Queuine tRNA-ribosyltransferase modifies tRNAs for asparagine.</p> <p>Comment: aspartic acid, histidine and tyrosine with queuine.</p> <p>Comment: It catalyses the exchange of guanine-34 at the wobble position with</p> <p>Comment: 7-aminomethyl-7-deazaguanine, and the addition of a cyclopentenediol</p> <p>Comment: moiety to 7-aminomethyl-7-deazaguanine-34 tRNA; giving a hypermodified</p> <p>Comment: base queuine in the wobble position [1,2].</p> <p>Comment: The aligned region contains a zinc binding motif C-x-C-x2-C-x29-H,</p> <p>Comment: and important tRNA and 7-aminomethyl-7-deazaguanine binding residues [1].</p> <p>Number of members: 24</p>
Thi4	Thi4 family	<p>Accession number: PF01946</p> <p>Definition: Thi4 family</p> <p>Author: Enright A, Ouzounis C, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Enright A</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 526.80 526 80</p> <p>Noise cutoffs: -105.00 -105.00</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 95050223</p> <p>Reference Title: Cloning, nucleotide sequence, and regulation of</p> <p>Reference Title: Schizosaccharomyces pombe thi4, a thiamine biosynthetic</p> <p>Reference Title: gene.</p> <p>Reference Author: Zur Linden A, Schweingruber ME,</p> <p>Reference Location: J Bacteriol 1994;176:6631-6635.</p> <p>Database Reference: INTERPRO; IPR002922;</p> <p>Comment: This family includes Swiss:P32318 a putative thiamine biosynthetic</p> <p>Comment: enzyme.</p> <p>Number of members: 14</p>

991

ThiC		ThiC family	<p>Accession number: PF01964 Definition: ThiC family Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: 25 25 Trusted cutoffs: 1047.20 1047.20 Noise cutoffs: -338.20 -338.20 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 93163063 Reference Title: Structural genes for thiamine biosynthetic enzymes (thiCEFGH) in <i>Escherichia coli</i> K-12. Reference Author: Vander Horn PB, Backstrom AD, Stewart V, Begley TP; Reference Location: J Bacteriol 1993;175:982-992. Reference Number: [2] Reference Medline: 99311269 Reference Title: Thiamin biosynthesis in prokaryotes. Reference Author: Begley TP, Downs DM, Ealick SE, McLafferty FW, Van Loon AP, Reference Author: Taylor S, Campobasso N, Chiu HJ, Kinsland C, Reddick JJ, Xi Reference Author: J; Reference Location: Arch Microbiol 1999;171:293-300. Reference Number: [3] Reference Medline: 97284509 Reference Title: Characterization of the <i>Bacillus subtilis</i> thiC operon involved in thiamine biosynthesis. Reference Author: Zhang Y, Taylor SV, Chiu HJ, Begley TP, Reference Location: J Bacteriol 1997;179:3030-3035. Database Reference: INTERPRO IPR002817, Comment: ThiC is found within the thiamine biosynthesis operon. ThiC is Comment: involved in pyrimidine biosynthesis [2]. Comment: ThiC catalyzes the substitution of the pyrophosphate of Comment: 2-methyl-4-amino-5-hydroxymethylpyrimidine pyrophosphate by Comment: 4-methyl-5-(beta-hydroxyethyl)thiazole phosphate to yield thiamine Comment: phosphate [3]. Number of members: 12</p>
ThiJ		ThiJ/Pfpl family	<p>Accession number: PF01965 Definition: ThiJ/Pfpl family Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: -40.2 -40.2 Trusted cutoffs: -40.20 -40.20 Noise cutoffs: -47.00 -47.00 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97039868 Reference Title: The thiJ locus and its relation to phosphorylation of hydroxymethylpyrimidine in <i>Escherichia coli</i>. Reference Author: Mizote T, Tsuda M, Nakazawa T, Nakayama H; Reference Location: Microbiology 1996;142:2969-2974 Reference Number: [2] Reference Medline: 96196168 Reference Title: Sequence, expression in <i>Escherichia coli</i>, and analysis of the gene encoding a novel intracellular protease (Pfpl) from the hyperthermophilic archaeon <i>Pyrococcus furiosus</i>. Reference Author: Halio SB, Blumentals II, Short SA, Merrill BM, Kelly RM; Reference Location: J Bacteriol 1996;178:2605-2612. Database Reference: INTERPRO; IPR002818, Database reference: PFAMB; PB002774; Database reference: PFAMB; PB007213; Database reference: PFAMB; PB041784; Comment: This family includes ThiJ a thiamine biosynthesis Comment: enzyme [1] that catalyses the phosphorylation of Comment: hydroxymethylpyrimidine (HMP) to HMP monophosphate</p>

			<p>EC:2.7.1.49. Comment: The family also includes a the protease Pfpl Swiss:Q51732 [2]. Number of members: 34</p>
Thr_dehydrat _C		C-terminal domain of Threonine dehydratase	<p>Accession number: PF00585 Definition: C-terminal domain of Threonine dehydratase Previous Pfam IDs Thr_dehydratase_C, Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Bateman A Gathering cutoffs: 25 25 Trusted cutoffs: 99.90 51.30 Noise cutoffs: -1.10 -1 10 HMM build command line: hmmbuild HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98230745 Reference Title: Structure and control of pyridoxal phosphate dependent Reference Title: allosteric threonine deaminase. Reference Author: Gallagher DT, Gilliland GL, Xiao G, Zondio J, Fisher KE, Reference Author Chinchilla D, Eisenstein E: Reference Location: Structure 1998;6:465-475. Database Reference: SCOP, 1tdj; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR001721, Database Reference: PDB, 1tdj ; 424; 512. Database Reference: PDB; 1tdj ; 329; 419; Comment: -!- Threonine dehydratases PALP all contain a carboxy Comment: terminal region. This region may have a regulatory role. Comment: Some members contain two copies of this region. Number of members: 30</p>
thymidylat_sy nt	PDOC00086	Thymidylate synthase active site	<p>Thymidylate synthase (EC 2.1.1.45) [1,2] catalyzes the reductive methylation of dUMP to dTMP with concomitant conversion of 5,10- methylenetetrahydrofolate to dihydrofolate. Thymidylate synthase plays an essential role in DNA synthesis and is an important target for certain chemotherapeutic drugs.</p> <p>Thymidylate synthase is an enzyme of about 30 to 35 Kd in most species except in protozoan and plants where it exists as a bifunctional enzyme that includes a dihydrofolate reductase domain.</p> <p>A cysteine residue is involved in the catalytic mechanism (it covalently binds the 5,6-dihydro-dUMP intermediate). The sequence around the active site of this enzyme is conserved from phages to vertebrates.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern R-x(2)-[LIVM]-x(3)-[FW]-[QN]-x(8,9)-[LV]-x-P-C-[HAVM]- x(3)-[QMT]-[FYW]-x-[LV] [C is the active site residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update November 1997 / Pattern and text revised. References [1] Benkovic S.J. Annu. Rev. Biochem. 49:227-251(1980). [2] Ross P., O'Gara F., Condon S. Appl. Environ. Microbiol. 56:2156-2163(1990).</p>
Top6A		Type II DNA topoisomeras e	<p>Accession number: PF01962 Definition: Type II DNA topoisomerase Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: -99 -99 Trusted cutoffs: -40.40 -40.40</p>

993

			<p>Noise cutoffs: -158.40 -158.40 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmlcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97238688 Reference Title: An atypical topoisomerase II from Archaea with implications Reference Title: for meiotic recombination [see comments] Reference Author: Bergerat A, de Massy B, Gadelle D, Varoutas PC, Nicolas A, Reference Author: Forterre P; Reference Location: Nature 1997;386:414-417. Database Reference: SCOP, 1d3y; fa: [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO, IPR002815, Database Reference: PDB; 1d3y A; 77, 363; Database Reference: PDB; 1d3y B; 77, 363, Comment: Members of this family are the A subunit from type II DNA Comment: topoisomerases Type II DNA topoisomerases catalyse Comment: the relaxation Comment: of DNA supercoiling by causing transient double strand Comment: breaks Comment: The family includes topoisomerase VI subunit A from Comment: archaeobacteria Comment: Swiss:Q57815 EC:5.99 1.3 and SPO11 from yeast Comment: Swiss:P23179. Comment: A conserved tyrosine is thought to be involved in breaking Comment: the Comment: double stranded DNA [1]. Number of members: 9</p>
Topoisom_ba c	PDOC00333	Prokaryotic DNA topoisomerases e I active site	<p>DNA topoisomerase I (EC 5.99 1 2) [1,2,3,4,E1] is one of the two types of enzyme that catalyze the interconversion of topological DNA isomers. Type I topoisomerases act by catalyzing the transient breakage of DNA, one strand at a time, and the subsequent rejoining of the strands. When a prokaryotic type I topoisomerase breaks a DNA backbone bond, it simultaneously forms a protein-DNA link where the hydroxyl group of a tyrosine residue is joined to a 5'-phosphate on DNA, at one end of the enzyme-severed DNA strand.</p> <p>Prokaryotic organisms, such as <i>Escherichia coli</i>, have two type I topoisomerase isozymes: topoisomerase I (gene <i>topA</i>) and topoisomerase III (gene <i>topB</i>). Eukaryotes also contain homologs of prokaryotic topoisomerase III</p> <p>There are a number of conserved residues in the region around the active site tyrosine; we used this region as a signature pattern</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [EQ]-x-L-Y-[DEQST]-x(3.12)-[LIV]-[ST]-Y-x-R-[ST]-[DEQS] [The second Y is the active site tyrosine] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update December 1999 / Pattern and text revised References [1] Sternglanz R Curr Opin. Cell Biol. 1:533-535(1990)</p> <p>[2] Sharma A., Mondragon A. Curr. Opin. Struct. Biol. 5:39-47(1995).</p> <p>[3] Bjornsti M.-A. Curr. Opin. Struct. Biol. 1:99-103(1991).</p> <p>[4] Roca J. Trends Biochem Sci. 20:156-160(1995).</p> <p>[E1]</p>

			http://ellington.pharm.arizona.edu/~bear/top/topo.html
toxin_3		long chain scorpion toxins	<p>Accession number: PF00537</p> <p>Definition: long chain scorpion toxins</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Manual</p> <p>Source of seed members: Arne Elofsson.</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 59 50 59 50</p> <p>Noise cutoffs: -3.80 -3.80</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Database Reference: SCOP; 2sn3; fa; [SCOP-USA][CATH-PDBSUM]</p> <p>Database Reference: INTERPRO; IPR002061;</p> <p>Comment: -I- Scorpion toxins bind to sodium channels and inhibit the activation</p> <p>Comment: mechanisms of the channels. thereby blocking neuronal transmission.</p> <p>Number of members: 77</p>
Translin		Translin family	<p>Accession number: PF01997</p> <p>Definition: Translin family</p> <p>Previous Pfam IDs: DUF130;</p> <p>Author: Enright A, Ouzounis C, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Enright A</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 299.50 299 50</p> <p>Noise cutoffs: -72.40 -72 40</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 97165975</p> <p>Reference Title: Isolation and characterization of a cDNA encoding a</p> <p>Reference Title: Translin-like protein, TRAX.</p> <p>Reference Author: Aoki K, Ishida R, Kasai M;</p> <p>Reference Location: FEBS Lett 1997;401:109-112.</p> <p>Database Reference: INTERPRO; IPR002848,</p> <p>Comment: Members of this family include Translin Swiss:Q15631 that interacts</p> <p>Comment: with DNA and forms a ring around the DNA. This family also includes</p> <p>Comment: Swiss:Q99598, that was found to interact with translin with yeast</p> <p>Comment: two-hybrid screen [1].</p> <p>Number of members: 10</p>
Transposase_19		Transposase 19	<p>Members of this family are capable of in vitro and/or in vivo insertion of a donor polynucleotide into a target polynucleotide. Such biological activity is useful for inserting DNA into host genome, for example, for cloning purposes to generate a desired vector in vitro.</p>
Transthyretin	PDOC00617	Transthyretin signatures	<p>Transthyretin (prealbumin) [1] is a thyroid hormone-binding protein that seems to transport thyroxine (T4) from the bloodstream to the brain. It is a protein of about 130 amino acids that assembles as a homotetramer and forms an internal channel that binds thyroxine. Transthyretin is mainly synthesized in the brain choroid plexus. In humans, variants of the protein are associated with distinct forms of amyloidosis</p> <p>The sequence of transthyretin is highly conserved in vertebrates. A number of uncharacterized proteins also belong to this family:</p> <ul style="list-style-type: none"> - Escherichia coli hypothetical protein yedX - Bacillus subtilis hypothetical protein yunM. - Caenorhabditis elegans hypothetical protein R09H10.3. - Caenorhabditis elegans hypothetical protein ZK697.8. <p>We selected two regions as signature patterns. The first located in the N-terminal extremity starts with a lysine known to be involved in binding T4. The second pattern is located in the C-terminal extremity.</p> <p>Description of pattern(s) and/or profile(s)</p>

			<p>Consensus pattern [KH]-[IV]-L-[DN]-x(3)-G-x-P-A-x(2)-[IV]-x-[IV] [The K binds thyroxine] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern Y-[TH]-[IV]-[AP]-x(2)-L-S-[PQ]-[FYW]-[GS]-[FY]-[QS] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE Last update July 1999 / Patterns and text revised. References [1] Schreiber G., Richardson S.J. Comp. Biochem Physiol 116B:137-160(1997).</p>
TRM		N2,N2-dimethylguanosine tRNA methyltransferase	<p>Accession number: PF02005 Definition: N2,N2-dimethylguanosine tRNA methyltransferase Author: Enright A, Ouzounis C, Bateman A Alignment method of seed: Clustalw Source of seed members: Enright A Gathering cutoffs: 25 25 Trusted cutoffs: 664.60 664.60 Noise cutoffs: -259.50 -259 50 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 98352211 Reference Title: The tRNA(guanine-26,N2-N2) methyltransferase (Trm1) from Reference Title: the hyperthermophilic archaeon Pyrococcus furiosus Reference Title: cloning, sequencing of the gene and its expression in Reference Title: Escherichia coli. Reference Author: Constantinesco F, Benachenhou N, Motorin Y, Grosjean H; Reference Location: Nucleic Acids Res 1998;26:3753-3761 Reference Number: [2] Reference Medline: 87260951 Reference Title: Amino-terminal extension generated from an upstream AUG Reference Title: codon is not required for mitochondrial import of yeast Reference Title: N2,N2-dimethylguanosine- specific tRNA methyltransferase. Reference Author: Ellis SR, Hopper AK, Martin NC; Reference Location: Proc Natl Acad Sci U S A 1987;84:5172-5176. Database Reference: INTERPRO; IPR002905, Database reference: PFAMB; PB041661; Comment: This enzyme EC:2.1.1.32 used S-AdoMet to methylate tRNA. Comment: The TRM1 gene of Saccharomyces cerevisiae is necessary for Comment: the N2,N2-dimethylguanosine modification of both mitochondrial Comment: and cytoplasmic tRNAs [1]. The enzyme is found in both Comment: eukaryotes and archaeobacteria [2] Number of members: 10</p>
tRNA bind		Putative tRNA binding domain	<p>Accession number: PF01588 Definition: Putative tRNA binding domain Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_482 (release 4 1) Gathering cutoffs: 20 20 Trusted cutoffs: 22 30 22.30 Noise cutoffs: 18.20 18.20 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97306356 Reference Title: Human tyrosyl-tRNA synthetase shares amino acid sequence Reference Title: homology with a putative cytokine. Reference Author: Kleeman TA, Wei D, Simpson KL, First EA; Reference Location: J Biol Chem 1997;272:14420-14425.</p>

			<p>Reference Number: [2] Reference Medline: 97050848 Reference Title: The yeast protein Arc1p binds to tRNA and functions as a cofactor for the methionyl- and glutamyl-tRNA synthetases. Reference Author: Simos G, Segref A, Fasiolo F, Hellmuth K, Shevchenko A, Reference Author: Mann M, Hurt EC; Reference Location: EMBO J 1996;15:5437-5448. Database Reference: SCOP; 1pys; fa; [SCOP-USA][CATH-PDBSUM] Database Reference: INTERPRO; IPR002547, Database Reference: PDB; 1b70 B; 153; 247; Database Reference: PDB; 1b7y B; 153; 247; Database Reference: PDB; 1ey B; 153; 247; Database Reference: PDB, 1pys B; 153; 247, Database reference: PFAMB; PB010015; Comment: This domain is found in prokaryotic methionyl-tRNA synthetases, Comment: prokaryotic phenylalanyl tRNA synthetases the yeast GU4 nucleic-binding Comment: protein (G4p1 or p42, ARC1) [2], human tyrosyl-tRNA synthetase [1], Comment: and endothelial-monocyte activating polypeptide II. Comment: G4p1 binds specifically to tRNA form a complex with methionyl-tRNA Comment: synthetases [2]. In human tyrosyl-tRNA synthetase this domain may direct Comment: tRNA to the active site of the enzyme [2] This domain may perform a Comment: common function in tRNA aminoacylation [1] Number of members: 46</p>
tRNA-synt_2d	PDOC00363	Aminoacyl-transfer RNA synthetases class-II signatures	<p>Aminoacyl-tRNA synthetases (EC 6.1.1 -) [1] are a group of enzymes which activate amino acids and transfer them to specific tRNA molecules as the first step in protein biosynthesis. In prokaryotic organisms there are at least twenty different types of aminoacyl-tRNA synthetases, one for each different amino acid. In eukaryotes there are generally two aminoacyl-tRNA synthetases for each different amino acid, one cytosolic form and a mitochondrial form. While all these enzymes have a common function, they are widely diverse in terms of subunit size and of quaternary structure.</p> <p>The synthetases specific for alanine, asparagine, aspartic acid, glycine, histidine, lysine, phenylalanine, proline, serine, and threonine are referred to as class-II synthetases [2 to 6] and probably have a common folding pattern in their catalytic domain for the binding of ATP and amino acid which is different to the Rossmann fold observed for the class I synthetases [7].</p> <p>Class-II tRNA synthetases do not share a high degree of similarity, however at least three conserved regions are present [2,5,8]. We have derived signature patterns from two of these regions.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [FYH]-R-x-[DE]-x(4,12)-[RH]-x(3)-F-x(3)-[DE] Sequences known to belong to this class detected by the pattern the majority of class-II tRNA synthetases with the exception of those specific for alanine, glycine as well as bacterial histidine. Other sequence(s) detected in SWISS-PROT 43.</p> <p>Consensus pattern [GSTALVF]-{DENQHRKP}-[GSTA]-[LIVMF]-[DE]-R-[LIVMF]-x-[LIVMSTAG]-[LIVMFY] Sequences known to belong to this class detected by the pattern the majority of class-II tRNA synthetases with the exception of those specific for serine and proline. Other sequence(s) detected in SWISS-PROT 161. Expert(s) to contact by email Cusack S. cusack@embl-grenoble.fr</p> <p>Last update July 1998 / Text revised. References [1] Schimmel P.</p>

			<p>Annu. Rev. Biochem. 56:125-158(1987).</p> <p>[2] Delarue M , Moras D. BioEssays 15:675-687(1993)</p> <p>[3] Schimmel P. Trends Biochem. Sci. 16:1-3(1991).</p> <p>[4] Nagel G.M , Doolittle R.F. Proc. Natl. Acad. Sci. U.S.A. 88:8121-8125(1991).</p> <p>[5] Cusack S., Haertlein M., Leberman R. Nucleic Acids Res. 19:3489-3498(1991).</p> <p>[6] Cusack S. Biochimie 75:1077-1081(1993).</p> <p>[7] Cusack S., Berthet-Colominas C., Haertlein M , Nassar N., Leberman R Nature 347:249-255(1990)</p> <p>[8] Leveque F , Plateau P., Dessen P., Blanquet S. Nucleic Acids Res. 18:305-312(1990).</p>
trypsin	PDOC00124	Serine proteases, trypsin family, active sites	<p>The catalytic activity of the serine proteases from the trypsin family is provided by a charge relay system involving an aspartic acid residue hydrogen-bonded to a histidine, which itself is hydrogen-bonded to a serine. The sequences in the vicinity of the active site serine and histidine residues are well conserved in this family of proteases [1]. A partial list of proteases known to belong to the trypsin family is shown below.</p> <ul style="list-style-type: none"> - Acrosin. - Blood coagulation factors VII, IX, X, XI and XII, thrombin, plasminogen, and protein C. - Cathepsin G. - Chymotrypsins. - Complement components C1r, C1s, C2. and complement factors B, D and I. - Complement-activating component of RA-reactive factor. - Cytotoxic cell proteases (granzymes A to H). - Duodenase I. - Elastases 1, 2, 3A, 3B (protease E), leukocyte (medullasin) - Enterokinase (EC 3.4.21.9) (enteropeptidase). - Hepatocyte growth factor activator. - Hepsin - Glandular (tissue) kallikreins (including EGF-binding protein types A, B, and C, NGF-gamma chain, gamma-renin, prostate specific antigen (PSA) and tonin). - Plasma kallikrein. - Mast cell proteases (MCP) 1 (chymase) to 8. - Myeloblastin (proteinase 3) (Wegener's autoantigen). - Plasminogen activators (urokinase-type. and tissue-type). - Trypsins I, II, III, and IV. - Tryptases. - Snake venom proteases such as ancrod batroxobin, cerastobin, flavoxobin, and protein C activator. - Collagenase from common cattle grub and collagenolytic protease from Atlantic sand fiddler crab - Apolipoprotein(a). - Blood fluke cercarial protease. - Drosophila trypsin like proteases: alpha, easter, snake-locus. - Drosophila protease stubble (gene sb). - Major mite fecal allergen Der p III. <p>All the above proteins belong to family S1 in the classification of peptidases [2,E1] and originate from eukaryotic species. It should be noted that bacterial proteases that belong to family S2A are similar enough in the regions of the active site residues that they can be picked up by the same</p>

			<p>patterns. These proteases are listed below.</p> <ul style="list-style-type: none"> - Achromobacter lyticus protease I. - Lysobacter alpha-lytic protease. - Streptogrisin A and B (Streptomyces proteases A and B). - Streptomyces griseus glutamyl endopeptidase II. - Streptomyces fradiae proteases 1 and 2. <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVM]-[ST]-A-[STAG]-H-C [H is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for complement components C1r and C1s, pig plasminogen, bovine protein C, rodent urokinase, anrod, gyroxin and two insect trypsins. Other sequence(s) detected in SWISS-PROT 14.</p> <p>Consensus pattern [DNSTAGC]-[GSTAPIMVQH]-x(2)-G-[DE]-S-G-[GS]-[SAPHV]-[LIVMFYWH]-[LIVMFYSTANQH] [S is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for 18 different proteases which have lost the first conserved glycine. Other sequence(s) detected in SWISS-PROT H influenzae protease HAP which belongs to family S6 and 3 other proteins.</p> <p>Note if a protein includes both the serine and the histidine active site signatures. the probability of it being a trypsin family serine protease is 100%</p> <p>Last update November 1997 / Text revised References [1] Brenner S. Nature 334:528-530(1988).</p> <p>[2] Rawlings N D., Barrett A J. Meth. Enzymol. 244 19-61(1994).</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?peptidas.txt</p>
TYA		TYA transposon protein	<p>Accession number PF01021 Definition: TYA transposon protein Author: Bateman A Alignment method of seed. Clustalw Source of seed members: Pfam-B_90 (release 3.0) Gathering cutoffs: 15 15 Trusted cutoffs 18.00 18.00 Noise cutoffs: 13 70 13.70 HMM build command line: hmmbuild -f HMM SEED HMM build command line: hmmcalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97404699 Reference Title: Cryo-electron microscopy structure of yeast Ty retrotransposon virus-like particles. Reference Author: Palmer KJ, Tichelaar W, Myers N, Burns NR, Butcher SJ, Reference Author: Kingsman AJ, Fuller SD, Saibil HR; Reference Location: J Virol 1997,71:6863-6868. Database Reference INTERPRO; IPR001042; Comment: Ty are yeast transposons. A 5.7kb transcript codes for p3 a fusion protein of TYA and TYB. The TYA protein is analogous to the gag protein of retroviruses. Comment: TYA is cleaved to form 46kd protein which can form mature virion like particles [1]. Number of members: 62</p>
tyrosinase	PDOC00398	Tyrosinase signatures	<p>Tyrosinase (EC 1.14.18.1) [1] is a copper monooxygenases that catalyzes the hydroxylation of monophenols and the oxidation of o-diphenols to o-quinols. This enzyme, found in prokaryotes as well as in eukaryotes, is involved in the formation of pigments such as melanins and other polyphenolic compounds.</p> <p>Tyrosinase binds two copper ions (CuA and CuB). Each of the two copper ion has</p>

		<p>been shown [2] to be bound by three conserved histidines residues. The regions around these copper-binding ligands are well conserved and also shared by some hemocyanins, which are copper-containing oxygen carriers from the hemolymph of many molluscs and arthropods [3,4].</p> <p>At least two proteins related to tyrosinase are known to exist in mammals:</p> <ul style="list-style-type: none"> - TRP-1 (TYRP1) [5], which is responsible for the conversion of 5,6-dihydroxyindole-2-carboxylic acid (DHICA) to indole-5,6-quinone-2-carboxylic acid. - TRP-2 (TYRP2) [6], which is the melanogenic enzyme DOPAchrome tautomerase (EC 5.3.3.12) that catalyzes the conversion of DOPAchrome to DHICA. TRP-2 differs from tyrosinases and TRP-1 in that it binds two zinc ions instead of copper [7]. <p>Other proteins that belong to this family are:</p> <ul style="list-style-type: none"> - Plants polyphenol oxidases (PPO) (EC 1.10.3.1) which catalyze the oxidation of mono- and o-diphenols to o-diquinones [8]. - <i>Caenorhabditis elegans</i> hypothetical protein C02C2.1. <p>We have derived two signature patterns for tyrosinase and related proteins. The first one contains two of the histidines that bind CuA, and is located in the N-terminal section of tyrosinase. The second pattern contains a histidine that binds CuB, that pattern is located in the central section of the enzyme.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern H-x(4,5)-F-[LIVMFTP]-x-[FW]-H-R-x(2)-[LVM]-x(3)-E [The two H's are copper ligands] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern D-P-x-F-[LIVMFYW]-x(2)-H-x(3)-D [H is a copper ligand] Sequences known to belong to this class detected by the pattern ALL the tyrosinases as well as all the hemocyanins Other sequence(s) detected in SWISS-PROT NONE</p> <p>Last update December 1999 / Patterns and text revised.</p> <p>References</p> <p>[1] Lerch K. Prog Clin. Biol. Res. 256:85-98(1988).</p> <p>[2] Jackman M.P., Hajnal A, Lerch K. Biochem. J. 274:707-713(1991).</p> <p>[3] Linzen B. Naturwissenschaften 76:206-211(1989).</p> <p>[4] Lang W.H, van Holde K.E Proc. Natl. Acad. Sci. U.S.A 88:244-248(1991).</p> <p>[5] Kobayashi T., Urabe K., Winder A., Jimenez-Cervantes C., Imokawa G., Brewington T., Solano F., Garcia-Borrón J.C., Hearing V.J. EMBO J. 13:5818-5825(1994).</p> <p>[6] Jackson I.J., Chambers D.M., Tsukamoto K., Copeland N.G., Gilbert D.J., Jenkins N.A., Hearing V. EMBO J. 11:527-535(1992).</p> <p>[7] Solano F., Martinez-Liarte J.H., Jimenez-Cervantes C, Garcia-Borrón J.C., Lozano J.A.</p>
--	--	---

1000

			<p>Biochem. Biophys. Res. Commun. 204:1243-1250(1994).</p> <p>[8] Cary J.W , Lax A.R , Flurkey W H. Plant Mol. Biol. 20:245-253(1992)</p>
UbiA	PDOC00727	UbiA prenyltransferase family signature	<p>The following prenyltransferases are evolutionary related [1,2]:</p> <ul style="list-style-type: none"> - Bacterial 4-hydroxybenzoate octaprenyltransferase (gene ubiA). - Yeast mitochondrial para-hydroxybenzoate--polyprenyltransferase (gene COQ2). - Protoheme IX farnesyltransferase (heme O synthase) from yeast and mammals (gene COX10) and from bacteria (genes cyoE or ctaB). <p>These proteins probably contain seven transmembrane segments. The best conserved region is located in a loop between the second and third of these segments and we used it as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern N-x(3)-[DEH]-x(2)-[LIMF]-D-x(2)-[VM]-x-R-[ST]-x(2)-R-x(4)-G Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE. Last update December 1999 / Pattern and text revised. References [1] Melzer M., Heide L. Biochim Biophys. Acta 1212 93-102(1994). [2] Mogi T., Saiki K , Anraku Y. Mol. Microbiol. 14:391-398(1994).</p>
Ubie_methyltransferase	PDOC00911	ubiE/COQ5 methyltransferase family signatures	<p>The following methyltransferases have been shown [1] to share regions of similarities:</p> <ul style="list-style-type: none"> - Escherichia coli ubiE, which is involved in both ubiquinone and menaquinone biosynthesis and which catalyzes the S-adenosylmethionine dependent methylation of 2-polyprenyl-6-methoxy-1,4-benzoquinol into 2-polyprenyl-3-methyl-6-methoxy-1,4-benzoquinol and of demethylmenaquinol into menaquinol. - Yeast COQ5, a ubiquinone biosynthesis methyltransferase. - Bacillus subtilis spore germination protein C2 (gene: gercB or gerC2), a probable menaquinone biosynthesis methyltransferase. - Lactococcus lactis gerC2 homolog. - Caenorhabditis elegans hypothetical protein ZK652.9. - Leishmania donovani amastigote-specific protein A41 <p>These are hydrophilic proteins of about 30 Kd (except for ZK652.9 which is 65 Kd). They can be picked up in the database by the following patterns.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern Y-D-x-M-N-x(2)-[LIVM]-S-x(3)-H-x(2)-W Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern R-V-[LIVM]-K-[PV]-[GM]-G-x-[LIVMF]-x(2)-[LIVM]-E-x-S Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE. Last update December 1999 / Pattern and text revised References [1] Lee P.T., Hsu A.Y., Ha H.T , Clarke C.F. J. Bacteriol. 179:1748-1754(1997).</p>

1001

ubiquitin	PDOC00271	Ubiquitin domain signature and profile	<p>Ubiquitin [1,2,3] is a protein of seventy six amino acid residues, found in all eukaryotic cells and whose sequence is extremely well conserved from protozoan to vertebrates. It plays a key role in a variety of cellular processes, such as ATP-dependent selective degradation of cellular proteins, maintenance of chromatin structure, regulation of gene expression, stress response and ribosome biogenesis</p> <p>In most species, there are many genes coding for ubiquitin. However they can be classified into two classes. The first class produces polyubiquitin molecules consisting of exact head to tail repeats of ubiquitin. The number of repeats is variable (up to twelve in a <i>Xenopus</i> gene). In the majority of polyubiquitin precursors, there is a final amino-acid after the last repeat. The second class of genes produces precursor proteins consisting of a single copy of ubiquitin fused to a C-terminal extension protein (CEP). There are two types of CEP proteins and both seem to be ribosomal proteins.</p> <p>Ubiquitin is a globular protein, the last four C-terminal residues (Leu-Arg-Gly-Gly) extending from the compact structure to form a 'tail', important for its function. The latter is mediated by the covalent conjugation of ubiquitin to target proteins, by an isopeptide linkage between the C-terminal glycine and the epsilon amino group of lysine residues in the target proteins.</p> <p>There are a number of proteins which are evolutionary related to ubiquitin.</p> <ul style="list-style-type: none"> - Ubiquitin-like proteins from baculoviruses as well as in some strains of bovine viral diarrhea viruses (BVDV). These proteins are highly similar to their eukaryotic counterparts. - Mammalian protein GDX [4] GDX is composed of two domains, a N-terminal ubiquitin-like domain of 74 residues and a C-terminal domain of 83 residues with some similarity with the thyroglobulin hormonogenic site. - Mammalian protein FAU [5]. FAU is a fusion protein which consist of a N-terminal ubiquitin-like protein of 74 residues fused to ribosomal protein S30. - Mouse protein NEDD-8 [6], a ubiquitin-like protein of 81 residues. - Human protein BAT3, a large fusion protein of 1132 residues that contains a N-terminal ubiquitin-like domain. - <i>Caenorhabditis elegans</i> protein ubl-1 [7]. Ubl-1 is a fusion protein which consist of a N-terminal ubiquitin-like protein of 70 residues fused to ribosomal protein S27A. - Yeast DNA repair protein RAD23 [8]. RAD23 contains a N-terminal domain that seems to be distantly, yet significantly, related to ubiquitin. - Mammalian RAD23-related proteins RAD23A and RAD23B. - Mammalian BCL-2 binding athanogene-1 (BAG-1) BAG-1 is a protein of 274 residues that contains a central ubiquitin-like domain - Human spliceosome associated protein 114 (SAP 114 or SF3A120). - Yeast protein DSK2, a protein involved in spindle pole body duplication and which contains a N-terminal ubiquitin-like domain. - Human protein CKAP1/TFCB, <i>Schizosaccharomyces pombe</i> protein alp11 and <i>Caenorhabditis elegans</i> hypothetical protein F53F4.3. These proteins contain a N-terminal ubiquitin domain and a C-terminal CAP-Gly domain (see <PDOC00660>). - <i>Schizosaccharomyces pombe</i> hypothetical protein SpAC26A3.16. This protein contains a N-terminal ubiquitin domain. - Yeast protein SMT3. - Human ubiquitin-like proteins SMT3A and SMT3B. - Human ubiquitin-like protein SMT3C (also known as PIC1, Ubl1, Sumo-1; Gmp-1 or Sentrin). This protein is involved in targeting ranGAP1 to the nuclear pore complex protein ranBP2. - SMT3-like proteins in plants and <i>Caenorhabditis elegans</i>. <p>To identify ubiquitin and related proteins we have developed a pattern based on conserved positions in the central section of the sequence. A profile was also developed that spans the complete length of the ubiquitin domain</p>
-----------	-----------	--	---

1002

			<p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern K-x(2)-[LIVM]-x-[DESAK]-x(3)-[LIVM]-[PA]-x(3)-Q-x-[LIVM]-[LIVMC]-[LIVMFY]-x-G-x(4)-[DE]</p> <p>Sequences known to belong to this class detected by the pattern ALL, except for the RAD23 and SMT3 subfamilies. BAG-1 and SAP 114.</p> <p>Other sequence(s) detected in SWISS-PROT NONE</p> <p>Sequences known to belong to this class detected by the profile ALL</p> <p>Other sequence(s) detected in SWISS-PROT NONE</p> <p>Note this documentation entry is linked to both a signature pattern and a profile. As the profile is much more sensitive than the pattern, you should use it if you have access to the necessary software tools to do so.</p> <p>Last update July 1998 / Text revised</p> <p>Bio/Technology 8:209-215(1990) References</p> <p>[1] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991).</p> <p>[2] Monia B.P., Ecker D.J., Croke S.T</p> <p>[3] Finley D., Varshavsky A. Trends Biochem. Sci. 10:343-347(1985).</p> <p>[4] Filippi M., Tribioli C., Toniolo D. Genomics 7:453-457(1990)</p> <p>[5] Olvera J., Wool I.G. J. Biol. Chem. 268:17967-17974(1993).</p> <p>[6] Kumar S., Yoshida Y., Noda M. Biochem Biophys. Res. Commun. 195:393-399(1993)</p> <p>[7] Jones D., Candido E.P J. Biol. Chem. 268:19545-19551(1993).</p> <p>[8] Melnick L., Sherman F J. Mol. Biol. 233:372-388(1993).</p>
UPF0004	PDOC00984	Uncharacterized protein family UPF0004 signature	<p>The following uncharacterized proteins have been shown [1] to share regions of similarities:</p> <ul style="list-style-type: none"> - Escherichia coli hypothetical protein yliG. - Escherichia coli hypothetical protein yleA and HI0019, the corresponding Haemophilus influenzae protein. - Bacillus subtilis hypothetical protein yqeV. - Helicobacter pylori hypothetical protein HP0269. - Helicobacter pylori hypothetical protein HP0285. - Mycoplasma iowae hypothetical protein in 16S RNA 5' region. - Mycobacterium tuberculosis hypothetical protein Rv2733c. - Rickettsia prowazekii hypothetical protein RP416. - Rickettsia prowazekii hypothetical protein RP808. - Synechocystis strain PCC 6803 hypothetical protein slr0082. - Synechocystis strain PCC 6803 hypothetical protein slr0996. - Methanococcus jannaschii hypothetical protein MJ0865 - Methanococcus jannaschii hypothetical protein MJ0867. - Caenorhabditis elegans hypothetical protein F25B5.5. <p>The size of these proteins range from 47 to 61 Kd. They contain six conserved cysteines, three of which are clustered in a region that can be used as a signature pattern</p>

1003

			<p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVM]-x-[LIVMT]-x(2)-G-C-x(3)-C-[STAN]-[FY]-C-x-[LIVMT]-x(4)-G</p> <p>Sequences known to belong to this class detected by the pattern ALL.</p> <p>Other sequence(s) detected in SWISS-PROT 2.</p> <p>Last update December 1999 / Pattern and text revised.</p> <p>References [1] Bairoch A. Unpublished observations (1997)</p>
UPF0013		Uncharacterized membrane protein family UPF0013	<p>Accession number: PF01554</p> <p>Definition: Uncharacterized membrane protein family UPF0013</p> <p>Author: Bateman A</p> <p>Alignment method of seed Clustalw</p> <p>Source of seed members: Pfam-B_163 (release 4 0)</p> <p>Gathering cutoffs: -26 -26</p> <p>Trusted cutoffs: -16.10 -16 10</p> <p>Noise cutoffs: -36 70 -36.70</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Database Reference: URL: http://www.expasy.ch/cgi-bin/lists?upflist.txt,</p> <p>Database Reference: INTERPRO; IPR002528;</p> <p>Database reference: PFAMB, PB041103;</p> <p>Comment: These proteins are integral membrane proteins of unknown</p> <p>Comment: function.</p> <p>Number of members: 47</p>
UPF0019	PDOC00949	Uncharacterized protein family UPF0019 signature	<p>The following uncharacterized proteins have been shown [1,2] to be highly similar:</p> <ul style="list-style-type: none"> - Yeast protein SNZ1, which may be involved in growth arrest and cellular response to nutrient limitation. - Yeast chromosome VI hypothetical protein YFL059w. - Yeast chromosome XIV hypothetical protein YNL333w. - Fission yeast hypothetical protein SpAC29B12.04. - Hevea brasiliensis ethylene-inducible protein HEVER - Stellaria longipes hypothetical protein H47. - Bacillus subtilis hypothetical protein yaaD. - Haemophilus influenzae hypothetical protein HI1647. - Mycobacterium leprae hypothetical protein M1CL581.12c. - Mycobacterium tuberculosis hypothetical protein MtCY1A10.27. - Archaeoglobus fulgidus hypothetical protein AF0508. - Methanococcus jannaschii hypothetical protein MJ0677. - Methanococcus vannielii hypothetical protein in tRNA/5S rRNA gene cluster. - Methanobacterium thermoautotrophicum hypothetical protein Mth666. <p>These are hydrophilic proteins of about 32 Kd. They can be picked up in the database by the following pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern L-P-V-[VT]-[NQL]-F-[AT]-A-G-G-[LIV]-A-T-P-A-D-A-A-[LM]</p> <p>Sequences known to belong to this class detected by the pattern ALL</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Last update July 1998 / Pattern and text revised.</p> <p>References [1] Sivasubramaniam S., Vanniasingham V.M., Tan C.T., Chua N.H. Plant Mol. Biol. 29:173-178(1995).</p> <p>[2] Braun E.L., Fuge E.K., Padilla P.A., Werner-Washburne M. J. Bacteriol. 178:6865-6872(1996)</p>
UPF0047	PDOC01018	Uncharacterized protein family	<p>The following uncharacterized proteins have been shown [1] to be highly similar:</p>

1004

		UPF0047 signature	<ul style="list-style-type: none"> - Bacillus subtilis hypothetical protein yugU - Escherichia coli hypothetical protein ybQ - Mycobacterium tuberculosis hypothetical protein MtCY9C4.12 - Synechocystis strain PCC 6803 hypothetical protein sli1880. - Archaeoglobus fulgidus hypothetical protein AF2050. - Methanococcus jannaschii hypothetical protein MJ1081. - Methanobacterium thermoautotrophicum hypothetical protein MTH771 - Fission yeast hypothetical protein SpAC4A8.02c. <p>These are small proteins of 14 to 16 Kd. They can be picked up in the database by the following pattern. This pattern is located in the C-terminal part of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern S-X(2)-[LIV]-x-[LIV]-x(2)-G-x(4)-G-T-W-Q-x-[LIV] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE Last update July 1998 / First entry. References [1] Bairoch A. Unpublished observations (1998)</p>
UPF0052		Uncharacterized protein family UPF0052	Accession number: PF01933 Definition: Uncharacterised protein family UPF0052 Author: Enright A, Ouzounis C, Bateman A Alignment method of seed. Clustalw Source of seed members: Enright A Gathering cutoffs: 25 25 Trusted cutoffs: 263.90 263.90 Noise cutoffs: -134.40 -134.40 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Database Reference INTERPRO; IPR002882, Number of members: 12
UPF0057	PDOC01013	Uncharacterized protein family UPF0057 signature	<p>The following uncharacterized proteins have been shown [1] to be evolutionary related:</p> <ul style="list-style-type: none"> - Barley low-temperature induced protein blt101. - Lophorium elongatum salt-stress induced protein ES13. - Yeast hypothetical proteins YDL123w, YDR276c, YDR525Bw and YJL151c. - Caenorhabditis elegans hypothetical proteins F47B7.1, T23F2.3, T23F2.4, T23F2.5 and ZK632.10. - Escherichia coli hypothetical protein yqaE. - Synechocystis strain PCC 6803 hypothetical protein ssr1169. <p>These are small proteins of from 52 to 140 amino-acid residues that contains two transmembrane domains. As a signature pattern we selected a region that corresponds to the end of the first transmembrane helix.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIV]-x-[STA]-[LIVF](3)-P-P-[LIVA]-[GA]-[IV]-x(4)-[GKN] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE Last update July 1998 / First entry. References [1] Rudd K.E., Humphery-Smith I., Wasinger V C., Bairoch A. Electrophoresis 19 536-544(1998).</p>
UPF0066	PDOC01022	Uncharacterized protein family UPF0066 signature	<p>The following uncharacterized proteins have been shown [1] to be evolutionary related:</p> <ul style="list-style-type: none"> - Escherichia coli hypothetical protein yaeB and HI0510, the corresponding Haemophilus influenzae protein.

			<ul style="list-style-type: none"> - Agrobacterium tumefaciens Ti plasmid protein virR. - Pseudomonas aeruginosa protein rcsF. - Archaeoglobus fulgidus hypothetical protein AF0241. - Archaeoglobus fulgidus hypothetical protein AF0433. - Methanococcus jannaschii hypothetical protein MJ1583. - Methanobacterium thermoautotrophicum hypothetical protein MTH1797. <p>These are proteins of from 120 to 240 amino-acid residues (with the exception of AF0433 which is 366 residues long). As a signature pattern we selected a conserved region in the central part of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern G-[AV]-F-[STA]-x-R-[SA]-x(2)-R-P-N Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE Last update July 1999 / First entry. References [1] Bairoch A. Unpublished observations (1998)</p>
UPF0076	PDOC00838	Uncharacterized protein family UPF0076 signature	<p>The following uncharacterized proteins have been shown [1] to share regions of similarities:</p> <ul style="list-style-type: none"> - Goat antigen UK114, a human homolog and the rat corresponding protein which is known as perchloric acid soluble protein (PSP1). PSP1 [2] may inhibit an initiation stage of cell-free protein synthesis. - Mouse heat-responsive protein HRSP12. - Yeast chromosome V hypothetical protein YER057c - Yeast chromosome IX hypothetical protein YIL051c. - Caenorhabditis elegans hypothetical protein C23G10.2 - Escherichia coli hypothetical protein ycdK. - Escherichia coli hypothetical protein yhaR. - Escherichia coli hypothetical protein yigF and HI0719, the corresponding Haemophilus influenzae protein. - Escherichia coli hypothetical protein yoaB. - Bacillus subtilis hypothetical protein yabJ. - Haemophilus influenzae hypothetical protein HI1627. - Helicobacter pylori hypothetical protein HP0944. - Lactococcus lactis aldR. - Myxococcus xanthus dfrA. - Synechocystis strain PCC 6803 hypothetical protein slr0709. - Rhizobium strain NGR234 symbiotic plasmid hypothetical protein y4sK. - Pyrococcus horikoshii hypothetical protein PH0854 <p>These are small proteins of around 15 Kd whose sequence is highly conserved. As a signature pattern, we selected a well conserved region located in the C-terminal part of these proteins.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [PA]-[ASTPV]-R-[SACVF]-x-[LIVMFY]-x(2)-[GSAKR]-x-[LMVA]-x(5,8)-[LIVM]-E-[MI] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 4. Last update July 1999 / Pattern and text revised. References [1] Bairoch A Unpublished observations (1995).</p> <p>[2] Oka T., Tsuji H., Noda C., Sakai K., Hong Y.-M., Suzuki I., Munoz S., Natori Y. J. Biol. Chem. 270:30060-30067(1995).</p>

[2]
Oka T., Tsuji H., Noda C., Sakai K., Hong Y.-M., Suzuki I., Munoz S., Natori Y.
J. Biol. Chem. 270:30060-30067(1995).

1006

UPF0099		Domain of unknown function UPF0099	<p>Accession number: PF01981</p> <p>Definition: Domain of unknown function UPF0099</p> <p>Previous Pfam IDs: DUF119:</p> <p>Author: Enright A, Ouzounis C, Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Enright A</p> <p>Gathering cutoffs: 25 25</p> <p>Trusted cutoffs: 132.80 132.80</p> <p>Noise cutoffs: -35.70 -35 70</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>HMM build command line: hmmbuild -F HMM SEED</p> <p>Database Reference: INTERPRO, IPR002833;</p> <p>Comment: This domain has no known function</p> <p>Number of members: 10</p>
UQ_con	PDOC00163	Ubiquitin-conjugating enzymes active site	<p>Ubiquitin-conjugating enzymes (EC 6.3.2.19) (UBC or E2 enzymes) [1,2,3] catalyze the covalent attachment of ubiquitin to target proteins. An activated ubiquitin moiety is transferred from an ubiquitin-activating enzyme (E1) to E2 which later ligates ubiquitin directly to substrate proteins with or without the assistance of 'N-end' recognizing proteins (E3).</p> <p>In most species there are many forms of UBC (at least 9 in yeast) which are implicated in diverse cellular functions.</p> <p>A cysteine residue is required for ubiquitin-thiolester formation. There is a single conserved cysteine in UBC's and the region around that residue is conserved in the sequence of known UBC isozymes. We have used that region as a signature pattern.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [FYWLSP]-H-[PC]-[NH]-[LIV]-x(3,4)-G-x-[LIV]-C-[LIV]-x-[LIV] [C is the active site residue]</p> <p>Sequences known to belong to this class detected by the pattern ALL. except for yeast UBC6 (DOA2).</p> <p>Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Expert(s) to contact by email Jentsch S. jentsch@zmbh.uni-heidelberg.de</p> <p>Last update July 1998 / Text revised.</p> <p>References [1] Jentsch S., Seufert W., Sommer T., Reins H.-A. Trends Biochem. Sci. 15:195-198(1990).</p> <p>[2] Jentsch S., Seufert W., Hauser H.-P. Biochim. Biophys. Acta 1089:127-139(1991).</p> <p>[3] Hershko A. Trends Biochem. Sci. 16:265-268(1991).</p>
urease_gamma	PDOC00133	Urease signatures	<p>Urease (EC 3.5.1.5) is a nickel-binding enzyme that catalyzes the hydrolysis of urea to carbon dioxide and ammonia [1]. Historically, it was the first enzyme to be crystallized (in 1926). It is mainly found in plant seeds, microorganisms and invertebrates. In plants, urease is a hexamer of identical chains. In bacteria [2], it consists of either two or three different subunits (alpha, beta and gamma).</p> <p>Urease binds two nickel ions per subunit; four histidine, an aspartate and a carbamated-lysine serve as ligands to these metals; an additional histidine is involved in the catalytic mechanism [3].</p> <p>As signatures for this enzyme, we selected a region that contains two histidine that bind one of the nickel ions and the region of the active site histidine.</p>

1007

			<p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern T-[AY]-[GA]-[GAT]-[LIVM]-D-x-H-[LIVM]-H-x(3)-P [The two H's bind nickel] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE</p> <p>Consensus pattern [LIVM](2)-[CT]-H-[HN]-L-x(3)-[LIVM]-x(2)-D-[LIVM]-x-F-A [H is the active site residue] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE. Last update November 1997 / Patterns and text revised. References [1] Takishima K , Suga T., Mamiya G Eur. J. Biochem. 175 151-165(1988).</p> <p>[2] Mobley H L T., Husinger R.P Microbiol Rev 53:85-108(1989).</p> <p>[3] Jabri E., Carr M.B., Hausinger R.P., Karplus P.A Science 268.998-1004(1995)</p>
UreD		UreD urease accessory protein	<p>Accession number: PF01774 Definition: UreD urease accessory protein Author: Bashton M, Bateman A Alignment method of seed. Clustalw Source of seed members Pfam-B_1109 (release 4.2) Gathering cutoffs: 25 25 Trusted cutoffs. 186.00 186 00 Noise cutoffs: -42.60 -42 60 HMM build command line. hmmbuild -F HMM SEED HMM build command line. hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 97352660 Reference Title: Characterization of UreG. identification of a Reference Title: UreD-UreF-UreG complex, and evidence suggesting that a Reference Title. nucleotide-binding site in UreG is required for in vivo Reference Title: metallocenter assembly of Klebsiella aerogenes urease. Reference Author: Moncrief MB, Hausinger RP: Reference Location: J Bacteriol 1997;179.4081-4086. Reference Number: [2] Reference Medline: 96146510 Reference Title: Organization of Ureaplasma urealyticum urease gene cluster Reference Title: and expression in a suppressor strain of Escherichia coli. Reference Author: Neyrolles O, Ferns S, Behbahani N, Montagnier L, Blanchard Reference Author: A, Reference Location: J Bacteriol 1996;178:647-655. Reference Number: [3] Reference Medline: 94211837 Reference Title. In vitro activation of urease apoprotein and role of UreD as a chaperone required for nickel metallocenter assembly. Reference Author: Park IS, Carr MB, Hausinger RP; Reference Location: Proc Natl Acad Sci U S A 1994;91:3233-3237. Database Reference: INTERPRO; IPR002669; Comment: UreD is a urease accessory protein. Urease urease hydrolyses Comment: urea into ammonia and carbamic acid [2]. UreD is involved in Comment: activation of the urease enzyme via the UreD-UreF-UreG-urease complex Comment: [1] and is required for urease nickel metallocenter assembly [3]. Comment: See also UreF UreF, UreG HypB_UreG Number of members. 23</p>
UreF		UreF	Accession number: PF01730

1008

			<p>Definition: UreF Author: Bashton M, Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_2037 (release 4.1) Gathering cutoffs: -31 -31 Trusted cutoffs: -14 30 -14 30 Noise cutoffs: -49 30 -49 30 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 96404789 Reference Title: Purification and activation properties of UreD-UreF-urease Reference Title: apoprotein complexes. Reference Author: Moncrief MB, Hausinger RP. Reference Location: J Bacteriol 1996;178:5417-5421. Reference Number: [2] Reference Medline: 96146510 Reference Title: Organization of Ureaplasma urealyticum urease gene cluster Reference Title: and expression in a suppressor strain of Escherichia coli. Reference Author: Neyrolles O, Ferris S, Behbahani N, Montagnier L, Blanchard Reference Author: A: Reference Location: J Bacteriol 1996;178:647-655. Database Reference: INTERPRO; IPR002639, Comment: This family consists of the Urease accessory protein Comment: UreF. The urease enzyme (urea amidohydrolase) Comment: hydrolyses urea into ammonia and carbamic acid [2]. Comment: UreF is proposed to modulate the activation process of Comment: urease by eliminating the binding of nickel ions to Comment: noncarbamylated protein [1]. Number of members: 20</p>
XPG_N	PDOC00658	XPG protein signatures	<p>Xeroderma pigmentosum (XP) [1] is a human autosomal recessive disease. characterized by a high incidence of sunlight-induced skin cancer. People's skin cells with this condition are hypersensitive to ultraviolet light. due to defects in the incision step of DNA excision repair. There are a minimum of seven genetic complementation groups involved in this pathway: XP-A to XP-G. The defect in XP-G can be corrected by a 133 Kd nuclear protein called XPG (or XPGC) [2].</p> <p>XPG belongs to a family of proteins [2,3,4,5,6] that are composed of two main subsets</p> <ul style="list-style-type: none"> - Subset 1 to which belongs XPG. RAD2 from budding yeast and rad13 from fission yeast. RAD2 and XPG are single-stranded DNA endonucleases [7,8]. XPG makes the 3' incision in human DNA nucleotide excision repair [9]. - Subset 2, to which belongs mouse and human FEN-1, rad2 from fission yeast, and RAD27 from budding yeast. FEN-1 is a structure-specific endonuclease. <p>In addition to the proteins listed in the above groups. this family also includes:</p> <ul style="list-style-type: none"> - Fission yeast exo1, a 5'->3' double-stranded DNA exonuclease that could act in a pathway that corrects mismatched base pairs. - Yeast EXO1 (DHS1), a protein with probably the same function as exo1. - Yeast DIN7. <p>Sequence alignment of this family of proteins reveals that similarities are largely confined to two regions. The first is located at the N-terminal extremity (N-region) and corresponds to the first 95 to 105 amino acids. The second region is internal (I-region) and found towards the C-terminus; it spans about 140 residues and contains a highly conserved core of 27 amino acids that includes a conserved pentapeptide (E-A-[DE]-A-[QS]). It is possible that the conserved acidic residues are involved in the catalytic mechanism of DNA excision repair in XPG. The amino acids linking the N- and I-regions are not conserved; indeed, they are largely absent from proteins belonging to the second subset.</p>

1009

			<p>We have developed two signature patterns for these proteins. The first corresponds to the central part of the N-region the second to part of the I-region and includes the putative catalytic core pentapeptide.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [VI]-[KRE]-P-x-[FYIL]-V-F-D-G-x(2)-[PIL]-x-[LVC]-K Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Consensus pattern [GS]-[LIVM]-[PER]-[FYS]-[LIVM]-x-A-P-x-E-A-[DE]-[PAS]-[QS]-[CLM] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT NONE. Expert(s) to contact by email Clarkson S.G. clarkson@medecine.unige.ch</p> <p>Last update November 1997 / Patterns and text revised. References [1] Tanaka K., Wood R.D Trends Biochem. Sci. 19 83-86(1994).</p> <p>[2] Scherly D., Nospikel T., Corlet J., Ucla C., Barroch A., Clarkson S.G. Nature 363:182-185(1993).</p> <p>[3] Carr A.M., Sheldrick K.S., Murray J.M., Al-Harthy R., Watts F.Z., Lehmann A.R. Nucleic Acids Res. 21:1345-1349(1993).</p> <p>[4] Murray J.M., Tavassoli M., Al-Harthy R., Sheldrick K.S., Lehmann A.R., Carr A.M., Watts F.Z. Mol. Cell. Biol. 14:4878-4888(1994).</p> <p>[5] Harrington J.J., Lieber M.R. Genes Dev. 8:1344-1355(1994).</p> <p>[6] Szankasi P., Smith G.R. Science 267:1166-1169(1995).</p> <p>[7] Habraken Y., Sung P., Prakash L., Prakash S. Nature 366:365-368(1993).</p> <p>[8] O'Donovan A., Scherly D., Clarkson S.G., Wood R.D. J. Biol. Chem. 269:15965-15968(1994).</p> <p>[9] O'Donovan A., Davies A.A., Moggs J.G., West S.C., Wood R.D. Nature 371:432-435(1994).</p>
Y_phosphatase	PDOC00323	Tyrosine specific protein phosphatases signature and profiles	<p>Tyrosine specific protein phosphatases (EC 3.1.3.48) (PTPase) [1 to 5] are enzymes that catalyze the removal of a phosphate group attached to a tyrosine residue. These enzymes are very important in the control of cell growth, proliferation, differentiation and transformation. Multiple forms of PTPase have been characterized and can be classified into two categories: soluble PTPases and transmembrane receptor proteins that contain PTPase domain(s). The currently known PTPases are listed below:</p> <p>Soluble PTPases.</p> <ul style="list-style-type: none"> - PTPN1 (PTP-1B). - PTPN2 (T-cell PTPase; TC-PTP). - PTPN3 (H1) and PTPN4 (MEG), enzymes that contain an N-terminal band

4.1-

like domain (see <PDOC00566>) and could act at junctions between the membrane and cytoskeleton.

- PTPN5 (STEP).
- PTPN6 (PTP-1C; HCP; SHP) and PTPN11 (PTP-2C; SH-PTP3; Syp).

enzymes which

- contain two copies of the SH2 domain at its N-terminal extremity. The *Drosophila* protein corkscrew (gene *csw*) also belongs to this subgroup.
- PTPN7 (LC-PTP; Hematopoietic protein-tyrosine phosphatase; HePTP).
- PTPN8 (70Z-PEP).
- PTPN9 (MEG2).
- PTPN12 (PTP-G1; PTP-P19).

- Yeast PTP1.
- Yeast PTP2 which may be involved in the ubiquitin-mediated protein degradation pathway.
- Fission yeast *pyp1* and *pyp2* which play a role in inhibiting the onset of mitosis.
- Fission yeast *pyp3* which contributes to the dephosphorylation of *cdc2*.
- Yeast CDC14 which may be involved in chromosome segregation.
- *Yersinia* virulence plasmid PTPases (gene *yopH*)
- *Autographa californica* nuclear polyhedrosis virus 19 Kd PTPase

Dual specificity PTPases.

- DUSP1 (PTPN10. MAP kinase phosphatase-1; MKP-1); which dephosphorylates MAP kinase on both Thr-183 and Tyr-185
- DUSP2 (PAC-1), a nuclear enzyme that dephosphorylates MAP kinases ERK1 and ERK2 on both Thr and Tyr residues
- DUSP3 (VHR).
- DUSP4 (HVB2)
- DUSP5 (HVB3).
- DUSP6 (Pyst1; MKP-3)
- DUSP7 (Pyst2; MKP-X).
- Yeast MSG5, a PTPase that dephosphorylates MAP kinase FUS3.
- Yeast YVH1.
- *Vaccinia* virus H1 PTPase, a dual specificity phosphatase

Receptor PTPases.

Structurally, all known receptor PTPases, are made up of a variable length extracellular domain, followed by a transmembrane region and a C-terminal catalytic cytoplasmic domain. Some of the receptor PTPases contain fibronectin type III (FN-III) repeats, immunoglobulin-like domains, MAM domains or carbonic anhydrase-like domains in their extracellular region. The cytoplasmic region generally contains two copies of the PTPase domain. The first seems to have enzymatic activity, while the second is inactive but seems to affect substrate specificity of the first. In these domains, the catalytic cysteine is generally conserved but some other, presumably important, residues are not.

In the following table, the domain structure of known receptor PTPases is shown:

	Extracellular		Intracellular			
	Ig	FN-3	CAH	MAM	PTPase	
Leukocyte common antigen (LCA) (CD45)			0	2	0	2
Leukocyte antigen related (LAR)			3	8	0	2
<i>Drosophila</i> DLAR			3	9	0	2
<i>Drosophila</i> DPTP			2	2	0	2
PTP-alpha (LRP)			0	0	0	2
PTP-beta			0	16	0	1
PTP-gamma			0	1	1	2
PTP-delta			0	>7	0	2
PTP-epsilon			0	0	0	2
PTP-kappa			1	4	0	2
PTP-mu			1	4	0	2
PTP-zeta			0	1	1	2

PTPase domains consist of about 300 amino acids. There are two conserved

1011

			<p>cysteines, the second one has been shown to be absolutely required for activity Furthermore, a number of conserved residues in its immediate vicinity have also been shown to be important.</p> <p>We derived a signature pattern for PTPase domains centered on the active site cysteine.</p> <p>There are three profiles for PTPases, the first one spans the complete domain and is not specific to any subtype. The second profile is specific to dual-specificity PTPases and the third one to the PTP subfamily.</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [LIVMF]-H-C-x(2)-G-x(3)-[STC]-[STAGP]-x-[LIVMFY] [C is the active site residue] Sequences known to belong to this class detected by the pattern ALL, except for nine sequences. Other sequence(s) detected in SWISS-PROT 3.</p> <p>Sequences known to belong to this class detected by the 1st profile ALL. Other sequence(s) detected in SWISS-PROT 2.</p> <p>Sequences known to belong to this class detected by the 2nd profile ALL dual type PTPases. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Sequences known to belong to this class detected by the 3rd profile ALL PTP type PTPases. Other sequence(s) detected in SWISS-PROT NONE.</p> <p>Note the M-phase inducer phosphatases (cdc25-type phosphatase) are tyrosine- protein phosphatases that are not structurally related to the above PTPases</p> <p>Note this documentation entry is linked to both a signature pattern and to profiles As profiles are much more sensitive than the pattern, you should use them if you have access to the necessary software tools to do so.</p> <p>Last update July 1999 / Text revised</p> <p>References</p> <p>[1] Fischer E.H., Charbonneau H., Tonks N.K Science 253:401-406(1991).</p> <p>[2] Charbonneau H., Tonks N.K. Annu. Rev. Cell Biol. 8:463-493(1992).</p> <p>[3] Trowbridge I.S. J. Biol. Chem 266:23517-23520(1991).</p> <p>[4] Tonks N.K., Charbonneau H. Trends Biochem. Sci. 14:497-500(1989).</p> <p>[5] Hunter T. Cell 58:1013-1016(1989)</p>
Zein		Zein seed storage protein	<p>Accession number: PF01559 Definition: Zein seed storage protein Author: Bateman A Alignment method of seed: Clustalw Source of seed members: Pfam-B_181 (release 4 0) Gathering cutoffs: -21 -21 Trusted cutoffs: 4.60 4.60 Noise cutoffs: -46.60 -46.60 HMM build command line: hmmbuild -F HMM SEED HMM build command line: hmmscalibrate --seed 0 HMM Reference Number: [1] Reference Medline: 93197294</p>

1012

			<p>Reference Title: Studies of the zein-like alpha-prolamins based on an analysis of amino acid sequences: implications for their evolution and three-dimensional structure</p> <p>Reference Author: Garratt R, Oliva G, Caracelli I, Leite A, Arruda P;</p> <p>Reference Location: Proteins 1993;15:88-99.</p> <p>Database Reference: INTERPRO; IPR002530;</p> <p>Comment: Zeins are seed storage proteins. They are unusually rich in glutamine, proline, alanine, and leucine residues and their sequences show a series of tandem repeats [1].</p> <p>Number of members: 48</p>
zf-AN1		AN1-like Zinc finger	<p>Accession number: PF01428</p> <p>Definition: AN1-like Zinc finger</p> <p>Author: Bateman A, SMART</p> <p>Alignment method of seed: Manual</p> <p>Source of seed members: SMART</p> <p>Gathering cutoffs: 16 16</p> <p>Trusted cutoffs: 16.40 16.40</p> <p>Noise cutoffs: 7.30 7.30</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 93292985</p> <p>Reference Title: Two related localized mRNAs from <i>Xenopus laevis</i> encode ubiquitin-like fusion proteins.</p> <p>Reference Author: Linnen JM, Bailey CP, Weeks DL;</p> <p>Reference Location: Gene 1993;128 181-188.</p> <p>Database reference: SMART; ZnF_AN1;</p> <p>Database Reference: INTERPRO; IPR000058;</p> <p>Comment: Zinc finger at the C-terminus of An1 Swiss:Q91889, a ubiquitin-like protein in <i>Xenopus laevis</i>.</p> <p>Comment: The following pattern describes the zinc finger.</p> <p>Comment: C-X2-C-X(9-12)-C-X(1-2)-C-X4-C-X2-H-X5-H-X-C</p> <p>Comment: Where X can be any amino acid, and numbers in brackets indicate the number of residues.</p> <p>Number of members: 18</p>
zf-CONSTANS		CONSTANS family zinc finger	<p>Accession number: PF01760</p> <p>Definition: CONSTANS family zinc finger</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_1072 (release 4.2)</p> <p>Gathering cutoffs: 25 10</p> <p>Trusted cutoffs: 76 10 17 20</p> <p>Noise cutoffs: 9 70 9.70</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 95211836</p> <p>Reference Title: The CONSTANS gene of <i>Arabidopsis</i> promotes flowering and encodes a protein showing similarities to zinc finger transcription factors.</p> <p>Reference Author: Putterill J, Robson F, Lee K, Simon R, Coupland G;</p> <p>Reference Location: Cell 1995;80:847-857.</p> <p>Database Reference: INTERPRO; IPR002926;</p> <p>Number of members: 45</p>
zf-DHHC		DHHC zinc finger domain	<p>Accession number: PF01529</p> <p>Definition: DHHC zinc finger domain</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Clustalw</p> <p>Source of seed members: Pfam-B_945 (release 4.0)</p> <p>Gathering cutoffs: 22 22</p> <p>Trusted cutoffs: 22.40 22.40</p> <p>Noise cutoffs: -22.40 -22.40</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 99250263</p> <p>Reference Title: The drosophila STAM gene homolog is in a tight gene</p>

1013

			<p>Reference Title: cluster, and its expression correlates to that of the adjacent gene ial.</p> <p>Reference Author: Mesilaty-Gross S, Reich A, Motro B, Wides R;</p> <p>Reference Location: Gene 1999;231:173-186.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 97315340</p> <p>Reference Title: Variations of the C2H2 zinc finger motif in the yeast genome and classification of yeast zinc finger proteins.</p> <p>Reference Author: Bohm S, Frishman D, Mewes HW;</p> <p>Reference Location: Nucleic Acids Res 1997;25:2464-2469</p> <p>Reference Number: [3]</p> <p>Reference Medline: 99321009</p> <p>Reference Title: The DHHC domain: a new highly conserved cysteine-rich motif</p> <p>Reference Author: Putilina T, Wong P, Gentleman S,</p> <p>Reference Location: Mol Cell Biochem 1999;195:219-226</p> <p>Reference Number: [4]</p> <p>Reference Medline: 10490616</p> <p>Reference Title: Erf2, a Novel Gene Product That Affects the Localization and Palmitoylation of Ras2 in Saccharomyces cerevisiae.</p> <p>Reference Author: Bartels DJ, Mitchell DA, Dong X, Deschenes RJ;</p> <p>Reference Location: Mol Cell Biol 1999;19:6775-6787.</p> <p>Database Reference: INTERPRO, IPR001594;</p> <p>Comment: This domain is also known as NEW1 [2]. This domain is predicted to be a zinc binding domain. The function of this domain is unknown, but it has been predicted to be involved in protein-protein or protein-DNA interactions [3].</p> <p>Number of members: 34</p>
zf-MYND		MYND finger	<p>Accession number: PF01753</p> <p>Definition: MYND finger</p> <p>Author: Bateman A</p> <p>Alignment method of seed: Manual</p> <p>Source of seed members: Bateman A</p> <p>Gathering cutoffs: 11 11</p> <p>Trusted cutoffs: 17.30 17.30</p> <p>Noise cutoffs: 5.50 5.50</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmscalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96203118</p> <p>Reference Title: DEAF-1, a novel protein that binds an essential region in a</p> <p>Reference Title: Deformed response element.</p> <p>Reference Author: Gross CT, McGinnis W;</p> <p>Reference Location: EMBO J 1996;15:1961-1970.</p> <p>Reference Number: [2]</p> <p>Reference Medline: 98079069</p> <p>Reference Title: Molecular cloning, sequence analysis, expression, and tissue distribution of suppressin, a novel suppressor of cell cycle entry.</p> <p>Reference Author: LeBoeuf RD, Ban EM, Green MM, Stone AS, Propst SM, Blalock</p> <p>Reference Author: JE, Tauber JD;</p> <p>Reference Location: J Biol Chem 1998;273:361-368.</p> <p>Database Reference: INTERPRO; IPR002893,</p> <p>Number of members: 48</p>
Zn_carbOpept	PDOC00123	Zinc carboxypeptidases, zinc-binding regions signatures	<p>There are a number of different types of zinc-dependent carboxypeptidases (EC 3.4.17.-) [1,2]. All these enzymes seem to be structurally and functionally related. The enzymes that belong to this family are listed below.</p> <ul style="list-style-type: none"> - Carboxypeptidase A1 (EC 3.4.17.1), a pancreatic digestive enzyme that removes all C-terminal amino acids with the exception of Arg, Lys and Pro. - Carboxypeptidase A2 (EC 3.4.17.15), a pancreatic digestive enzyme with a specificity similar to that of carboxypeptidase A1, but with a preference for bulkier C-terminal residues. - Carboxypeptidase B (EC 3.4.17.2), also a pancreatic digestive enzyme, but that preferentially removes C-terminal Arg and Lys. - Carboxypeptidase N (EC 3.4.17.3) (also known as arginine carboxypeptidase), a plasma enzyme which protects the body from potent vasoactive and

1014

		<p>inflammatory peptides containing C-terminal Arg or Lys (such as kinins or anaphylatoxins) which are released into the circulation.</p> <ul style="list-style-type: none"> - Carboxypeptidase H (EC 3.4.17.10) (also known as enkephalin convertase or carboxypeptidase E), an enzyme located in secretory granules of pancreatic islets, adrenal gland, pituitary and brain. This enzyme removes residual C-terminal Arg or Lys remaining after initial endoprotease cleavage during prohormone processing. - Carboxypeptidase M (EC 3.4.17.12), a membrane bound Arg and Lys specific enzyme. <p>It is ideally situated to act on peptide hormones at local tissue sites where it could control their activity before or after interaction with specific plasma membrane receptors.</p> <ul style="list-style-type: none"> - Mast cell carboxypeptidase (EC 3.4.17.1), an enzyme with a specificity to carboxypeptidase A, but found in the secretory granules of mast cells. - <i>Streptomyces griseus</i> carboxypeptidase (Cpase SG) (EC 3.4.17.-) [3], which combines the specificities of mammalian carboxypeptidases A and B - <i>Thermoactinomyces vulgaris</i> carboxypeptidase T (EC 3.4.17.18) (CPT) [4], which also combines the specificities of carboxypeptidases A and B. - AEBP1 [5], a transcriptional repressor active in preadipocytes. AEBP1 seems to regulate transcription by cleavage of other transcriptional proteins. - Yeast hypothetical protein YHR132c. <p>All of these enzymes bind an atom of zinc. Three conserved residues are implicated in the binding of the zinc atom: two histidines and a glutamic acid. We have derived two signature patterns which contain these three zinc-ligands</p> <p>Description of pattern(s) and/or profile(s)</p> <p>Consensus pattern [PK]-x-[LIVMFY]-x-[LIVMFY]-x(4)-H-[STAG]-x-E-x-[LIVM]-[STAG]-x(6)-[LIVMFYTA] [H and E are zinc ligands] Sequences known to belong to this class detected by the pattern ALL Other sequence(s) detected in SWISS-PROT <i>Bacillus sphaericus</i> endopeptidase I which hydrolyses the gamma-D-Glu-(L)meso-diaminopimelic acid bond of spore cortex peptidoglycan [6] and which is possibly distantly related to zinc carboxypeptidases.</p> <p>Consensus pattern H-[STAG]-x(3)-[LIVME]-x(2)-[LIVMFYW]-P-[FYW] [H is a zinc ligand] Sequences known to belong to this class detected by the pattern ALL. Other sequence(s) detected in SWISS-PROT 40.</p> <p>Note if a protein includes both signatures, the probability of it being a eukaryotic zinc carboxypeptidase is 100%</p> <p>Note these proteins belong to families M14A/M14B in the classification of peptidases [7,E1]. Last update November 1995 / Patterns and text revised. References [1] Tan F., Chan S J., Steiner D.F., Schilling J W., Skidgel R.A. J. Biol. Chem. 264:13165-13170(1989). [2] Reynolds D.S., Stevens R.L., Gurley D.S., Lane W.S., Austen K F., Serafin W.E. J. Biol. Chem. 264:20094-20099(1989). [3] Narahashi Y. J. Biochem. 107:879-886(1990). [4] Teplyakov A., Polyakov K., Obmolova G., Strokopytov B., Kuranova I., Osterman A.L., Grishin N.V., Smulevitch S.V., Zagnitko O.P., Galperina O.V., Matz M.V., Stepanov V.M. Eur. J. Biochem. 208:281-288(1992). [5] He G.-P., Muise A., Li A.W., Ro H.-S.</p>
--	--	---

1015

			<p>Nature 378:92-96(1995).</p> <p>[6] Hourdou M.-L., Guinand M., Vacheron M.J., Michel G., Denoroy L., Duez C.M., Englebert S., Joris B., Weber G., Ghuysen J.-M Biochem. J. 292:563-570(1993).</p> <p>[7] Rawlings N D., Barrett A.J. Meth. Enzymol. 248:183-228(1995).</p> <p>[E1] http://www.expasy.ch/cgi-bin/lists?peptidas.txt</p>
ZZ		Zinc finger present in dystrophin, CBP/p300	<p>Accession number: PF00569</p> <p>Definition: Zinc finger present in dystrophin, CBP/p300</p> <p>Author: SMART</p> <p>Alignment method of seed. Manual</p> <p>Source of seed members: Alignment kindly provided by SMART</p> <p>Gathering cutoffs: 14 14</p> <p>Trusted cutoffs. 14 60 14.60</p> <p>Noise cutoffs: 10.90 10.90</p> <p>HMM build command line: hmmbuild HMM SEED</p> <p>HMM build command line: hmmcalibrate --seed 0 HMM</p> <p>Reference Number: [1]</p> <p>Reference Medline: 96402609</p> <p>Reference Title: ZZ and TAZ new putative zinc fingers in dystrophin and other proteins.</p> <p>Reference Author: Ponting CP, Blake DJ, Davies KE, Kendrick-Jones J. Winder</p> <p>Reference Author: SJ:</p> <p>Reference Location: Trends Biochem Sci 1996;21.11-13.</p> <p>Database Reference: EXPERT; Chris.Ponting@human-anatomy.oxford.ac.uk:</p> <p>Database Reference INTERPRO; IPR000433,</p> <p>Database reference: PFAMB; PB041629;</p> <p>Comment: ZZ in dystrophin binds calmodulin</p> <p>Comment: Putative zinc finger, binding not yet shown.</p> <p>Number of members: 87</p>

AA. Activities of Polypeptides Comprising Signal Peptides

Polypeptides comprising signal peptides are a family of proteins that are typically
5 targeted to (1) a particular organelle or intracellular compartment, (2) interact with a
particular molecule or (3) for secretion outside of a host cell. Example of polypeptides
comprising signal peptides include, without limitation, secreted proteins, soluble proteins,
receptors, proteins retained in the ER, etc.

10 These proteins comprising signal peptides are useful to modulate ligand-receptor
interactions, cell-to-cell communication, signal transduction, intracellular communication,
and activities and/or chemical cascades that take part in an organism outside or within of any
particular cell.

15 One class of such proteins are soluble proteins which are transported out of the cell.
These proteins can act as ligands that bind to receptor to trigger signal transduction or to
permit communication between cells.

20 Another class is receptor proteins which also comprise a retention domain that lodges
the receptor protein in the membrane when the cell transports the receptor to the surface of
the cell. Like the soluble ligands, receptors can also modulate signal transduction and
communication between cells.

25 In addition the signal peptide itself can serve as a ligand for some receptors. An
example is the interaction of the ER targeting signal peptide with the signal recognition
particle (SRP). Here, the SRP binds to the signal peptide, halting translation, and the
resulting SRP complex then binds to docking proteins located on the surface of the ER,
prompting transfer of the protein into the ER.

30 A description of signal peptide residue composition is described below in Subsection
IV.C.1.

III. Methods of Modulating Polypeptide Production

It is contemplated that polynucleotides of the invention can be incorporated into a host cell or in-vitro system to modulate polypeptide production. For instance, the SDFs prepared as described herein can be used to prepare expression cassettes useful in a number of techniques for suppressing or enhancing expression.

An example are polynucleotides comprising sequences to be transcribed, such as coding sequences, of the present invention can be inserted into nucleic acid constructs to modulate polypeptide production. Typically, such sequences to be transcribed are heterologous to at least one element of the nucleic acid construct to generate a chimeric gene or construct.

Another example of useful polynucleotides are nucleic acid molecules comprising regulatory sequences of the present invention. Chimeric genes or constructs can be generated when the regulatory sequences of the invention linked to heterologous sequences in a vector construct. Within the scope of invention are such chimeric gene and/or constructs.

Also within the scope of the invention are nucleic acid molecules, whereof at least a part or fragment of these DNA molecules are presented in Tables 1 and 2 of the present application, and wherein the coding sequence is under the control of its own promoter and/or its own regulatory elements. Such molecules are useful for transforming the genome of a host cell or an organism regenerated from said host cell for modulating polypeptide production.

Additionally, a vector capable of producing the oligonucleotide can be inserted into the host cell to deliver the oligonucleotide.

More detailed description of components to be included in vector constructs are described both above and below.

Whether the chimeric vectors or native nucleic acids are utilized, such polynucleotides can be incorporated into a host cell to modulate polypeptide production. Native genes and/or nucleic acid molecules can be effective when exogenous to the host cell.

Methods of modulating polypeptide expression includes, without limitation:

Suppression methods, such as

Antisense

Ribozymes

Co-suppression

Insertion of Sequences into the Gene to be Modulated

Regulatory Sequence Modulation.

as well as Methods for Enhancing Production, such as
Insertion of Exogenous Sequences; and
Regulatory Sequence Modulation.

III.A. Suppression

Expression cassettes of the invention can be used to suppress expression of endogenous genes which comprise the SDF sequence. Inhibiting expression can be useful, for instance, to tailor the ripening characteristics of a fruit (Oeller et al., *Science* 254:437 (1991)) or to influence seed size (WO98/07842) or to provoke cell ablation (Mariani et al., *Nature* 357: 384-387 (1992)).

As described above, a number of methods can be used to inhibit gene expression in plants, such as antisense, ribozyme, introduction of exogenous genes into a host cell, insertion of a polynucleotide sequence into the coding sequence and/or the promoter of the endogenous gene of interest, and the like.

III.A.1. Antisense

An expression cassette as described above can be transformed into host cell or plant to produce an antisense strand of RNA. For plant cells, antisense RNA inhibits gene expression by preventing the accumulation of mRNA which encodes the enzyme of interest, *see*, e.g., Sheehy et al., *Proc. Nat. Acad. Sci. USA*, 85:8805 (1988), and Hiatt et al., U.S. Patent No. 4,801,340.

III.A.2. Ribozymes

Similarly, ribozyme constructs can be transformed into a plant to cleave mRNA and down-regulate translation.

III.A.3. Co-Suppression

Another method of suppression is by introducing an exogenous copy of the gene to be suppressed. Introduction of expression cassettes in which a nucleic acid is configured in the sense orientation with respect to the promoter has been shown to prevent the accumulation of mRNA. A detailed description of this method is described above.

III.A.4. Insertion of Sequences into the Gene to be Modulated

Yet another means of suppressing gene expression is to insert a polynucleotide into the gene of interest to disrupt transcription or translation of the gene.

Homologous recombination could be used to target a polynucleotide insert to a gene using the Cre-Lox system (A.C. Vergunst et al., *Nucleic Acids Res.* 26:2729 (1998), A.C. Vergunst et al., *Plant Mol. Biol.* 38:393 (1998), H. Albert et al., *Plant J.* 7:649 (1995)).

In addition, random insertion of polynucleotides into a host cell genome can also be used to disrupt the gene of interest. Azpiroz-Leehan et al., *Trends in Genetics* 13:152 (1997). In this method, screening for clones from a library containing random insertions is preferred for identifying those that have polynucleotides inserted into the gene of interest. Such screening can be performed using probes and/or primers described above based on sequences from Tables 1 and 2, fragments thereof, and substantially similar sequence thereto. The screening can also be performed by selecting clones or any transgenic plants having a desired phenotype.

III.A.5. Regulatory Sequence Modulation

The SDFs described in Tables 1 and 2, and fragments thereof are examples of nucleotides of the invention that contain regulatory sequences that can be used to suppress or inactivate transcription and/or translation from a gene of interest as discussed in I.C.5.

III.A.6. Genes Comprising Dominant-Negative Mutations

When suppression of production of the endogenous, native protein is desired it is often helpful to express a gene comprising a dominant negative mutation. Production of protein variants produced from genes comprising dominant negative mutations is a useful tool for research. Genes comprising dominant negative mutations can produce a variant polypeptide which is capable of competing with the native polypeptide, but which does not produce the native result. Consequently, over expression of genes comprising these mutations can titrate out an undesired activity of the native protein. For example, The product from a gene comprising a dominant negative mutation of a receptor can be used to constitutively activate or suppress a signal transduction cascade, allowing examination of the phenotype and thus the trait(s) controlled by that receptor and pathway. Alternatively, the protein arising from the gene comprising a dominant-negative mutation can be an inactive enzyme still capable

1020

of binding to the same substrate as the native protein and therefore competes with such native protein.

Products from genes comprising dominant-negative mutations can also act upon the native protein itself to prevent activity. For example, the native protein may be active only as a homo-multimer or as one subunit of a hetero-multimer. Incorporation of an inactive subunit into the multimer with native subunit(s) can inhibit activity.

Thus, gene function can be modulated in host cells of interest by insertion into these cells vector constructs comprising a gene comprising a dominant-negative mutation.

III.B. Enhanced Expression

Enhanced expression of a gene of interest in a host cell can be accomplished by either (1) insertion of an exogenous gene; or (2) promoter modulation.

III.B.1. Insertion of an Exogenous Gene

Insertion of an expression construct encoding an exogenous gene can boost the number of gene copies expressed in a host cell.

Such expression constructs can comprise genes that either encode the native protein that is of interest or that encode a variant that exhibits enhanced activity as compared to the native protein. Such genes encoding proteins of interest can be constructed from the sequences from Tables 1 and 2, fragments thereof, and substantially similar sequence thereto.

Such an exogenous gene can include either a constitutive promoter permitting expression in any cell in a host organism or a promoter that directs transcription only in particular cells or times during a host cell life cycle or in response to environmental stimuli.

III.B.2. Regulatory Sequence Modulation

The SDFs of Tables 1 and 2, and fragments thereof, contain regulatory sequences that can be used to enhance expression of a gene of interest. For example, some of these sequences contain useful enhancer elements. In some cases, duplication of enhancer elements or insertion of exogenous enhancer elements will increase expression of a desired gene from a particular promoter. As other examples, all II promoters require binding of a regulatory protein to be activated, while some promoters may need a protein that signals a promoter binding protein to expose a polymerase binding site. In either case, over-production of such proteins

can be used to enhance expression of a gene of interest by increasing the activation time of the promoter.

Such regulatory proteins are encoded by some of the sequences in Tables 1 and 2, fragments thereof, and substantially similar sequences thereto.

Coding sequences for these proteins can be constructed as described above.

IV. Gene Constructs and Vector Construction

To use isolated SDFs of the present invention or a combination of them or parts and/or mutants and/or fusions of said SDFs in the above techniques, recombinant DNA vectors which comprise said SDFs and are suitable for transformation of cells, such as plant cells, are usually prepared. The SDF construct can be made using standard recombinant DNA techniques (Sambrook et al. 1989) and can be introduced to the species of interest by *Agrobacterium*-mediated transformation or by other means of transformation (e.g., particle gun bombardment) as referenced below.

The vector backbone can be any of those typical in the art such as plasmids, viruses, artificial chromosomes, BACs, YACs and PACs and vectors of the sort described by

- (a) **BAC:** Shizuya et al., Proc. Natl. Acad. Sci. USA 89: 8794-8797 (1992); Hamilton et al., Proc. Natl. Acad. Sci. USA 93: 9975-9979 (1996);
- (b) **YAC:** Burke et al., Science 236:806-812 (1987);.
- (c) **PAC:** Sternberg N. et al., Proc Natl Acad Sci U S A. Jan;87(1):103-7 (1990);
- (d) **Bacteria-Yeast Shuttle Vectors:** Bradshaw et al., Nucl Acids Res 23: 4850-4856 (1995);
- (e) **Lambda Phage Vectors:** Replacement Vector, e.g., Frischauf et al., J. Mol Biol 170: 827-842 (1983); or Insertion vector, e.g., Huynh et al., In: Glover NM (ed) DNA Cloning: A practical Approach, Vol.1 Oxford: IRL Press (1985);
- (f) **T-DNA gene fusion vectors :**Walden et al., Mol Cell Biol 1: 175-194 (1990); and
- (g) **Plasmid vectors:** Sambrook et al., infra.

Typically, a vector will comprise the exogenous gene, which in its turn comprises an SDF of the present invention to be introduced into the genome of a host cell, and which gene may be an antisense construct, a ribozyme construct chimera, or a coding sequence with

any desired transcriptional and/or translational regulatory sequences, such as promoters, UTRs, and 3' end termination sequences. Vectors of the invention can also include origins of replication, scaffold attachment regions (SARs), markers, homologous sequences, introns, etc.

A DNA sequence coding for the desired polypeptide, for example a cDNA sequence encoding a full length protein, will preferably be combined with transcriptional and translational initiation regulatory sequences which will direct the transcription of the sequence from the gene in the intended tissues of the transformed plant.

For example, for over-expression, a plant promoter fragment may be employed that will direct transcription of the gene in all tissues of a regenerated plant. Alternatively, the plant promoter may direct transcription of an SDF of the invention in a specific tissue (tissue-specific promoters) or may be otherwise under more precise environmental control (inducible promoters).

If proper polypeptide production is desired, a polyadenylation region at the 3'-end of the coding region is typically included. The polyadenylation region can be derived from the natural gene, from a variety of other plant genes, or from T-DNA.

The vector comprising the sequences from genes or SDF or the invention may comprise a marker gene that confers a selectable phenotype on plant cells. The vector can include promoter and coding sequence, for instance. For example, the marker may encode biocide resistance, particularly antibiotic resistance, such as resistance to kanamycin, G418, bleomycin, hygromycin, or herbicide resistance, such as resistance to chlorosulfuron or phosphinotricin.

IV.A. Coding Sequences

Generally, the sequence in the transformation vector and to be introduced into the genome of the host cell does not need to be absolutely identical to an SDF of the present invention. Also, it is not necessary for it to be full length, relative to either the primary transcription product or fully processed mRNA. Furthermore, the introduced sequence need not have the same intron or exon pattern as a native gene. Also, heterologous non-coding segments can be incorporated into the coding sequence without changing the desired amino acid sequence of the polypeptide to be produced.

IV.B. Promoters

As explained above, introducing an exogenous SDF from the same species or an orthologous SDF from another species can modulate the expression of a native gene corresponding to that SDF of interest. Such an SDF construct can be under the control of either a constitutive promoter or a highly regulated inducible promoter (*e.g.*, a copper inducible promoter). The promoter of interest can initially be either endogenous or heterologous to the species in question. When re-introduced into the genome of said species, such promoter becomes exogenous to said species. Over-expression of an SDF transgene can lead to co-suppression of the homologous endogeneous sequence thereby creating some alterations in the phenotypes of the transformed species as demonstrated by similar analysis of the chalcone synthase gene (Napoli et al., *Plant Cell* 2:279 (1990) and van der Krol et al., *Plant Cell* 2:291 (1990)). If an SDF is found to encode a protein with desirable characteristics, its over-production can be controlled so that its accumulation can be manipulated in an organ- or tissue-specific manner utilizing a promoter having such specificity.

Likewise, if the promoter of an SDF (or an SDF that includes a promoter) is found to be tissue-specific or developmentally regulated, such a promoter can be utilized to drive or facilitate the transcription of a specific gene of interest (*e.g.*, seed storage protein or root-specific protein). Thus, the level of accumulation of a particular protein can be manipulated or its spatial localization in an organ- or tissue- specific manner can be altered.

IV. C Signal Peptides

SDFs of the present invention containing signal peptides are indicated in Tables 1 and 2. In some cases it may be desirable for the protein encoded by an introduced exogenous or orthologous SDF to be targeted (1) to a particular organelle intracellular compartment, (2) to interact with a particular molecule such as a membrane molecule or (3) for secretion outside of the cell harboring the introduced SDF. This will be accomplished using a signal peptide.

Signal peptides direct protein targeting, are involved in ligand-receptor interactions and act in cell to cell communication. Many proteins, especially soluble proteins, contain a signal peptide that targets the protein to one of several different intracellular compartments. In plants, these compartments include, but are not limited to, the endoplasmic reticulum (ER), mitochondria, plastids (such as chloroplasts), the vacuole, the Golgi apparatus, protein storage vesicles (PSV) and, in general, membranes. Some signal peptide sequences are conserved, such as the Asn-Pro-Ile-Arg amino acid motif found in the N-terminal propeptide

signal that targets proteins to the vacuole (Marty (1999) *The Plant Cell* 11: 587-599). Other signal peptides do not have a consensus sequence *per se*, but are largely composed of hydrophobic amino acids, such as those signal peptides targeting proteins to the ER (Vitale and Denecke (1999) *The Plant Cell* 11: 615-628). Still others do not appear to contain either a consensus sequence or an identified common secondary sequence, for instance the chloroplast stromal targeting signal peptides (Keegstra and Cline (1999) *The Plant Cell* 11: 557-570). Furthermore, some targeting peptides are bipartite, directing proteins first to an organelle and then to a membrane within the organelle (e.g. within the thylakoid lumen of the chloroplast; see Keegstra and Cline (1999) *The Plant Cell* 11: 557-570). In addition to the diversity in sequence and secondary structure, placement of the signal peptide is also varied. Proteins destined for the vacuole, for example, have targeting signal peptides found at the N-terminus, at the C-terminus and at a surface location in mature, folded proteins. Signal peptides also serve as ligands for some receptors.

These characteristics of signal proteins can be used to more tightly control the phenotypic expression of introduced SDFs. In particular, associating the appropriate signal sequence with a specific SDF can allow sequestering of the protein in specific organelles (plastids, as an example), secretion outside of the cell, targeting interaction with particular receptors, etc. Hence, the inclusion of signal proteins in constructs involving the SDFs of the invention increases the range of manipulation of SDF phenotypic expression. The nucleotide sequence of the signal peptide can be isolated from characterized genes using common molecular biological techniques or can be synthesized *in vitro*.

In addition, the native signal peptide sequences, both amino acid and nucleotide, described in Tables 1 and 2 can be used to modulate polypeptide transport. Further variants of the native signal peptides described in Tables 1 and 2 are contemplated. Insertions, deletions, or substitutions can be made. Such variants will retain at least one of the functions of the native signal peptide as well as exhibiting some degree of sequence identity to the native sequence.

Also, fragments of the signal peptides of the invention are useful and can be fused with other signal peptides of interest to modulate transport of a polypeptide.

V. Transformation Techniques

A wide range of techniques for inserting exogenous polynucleotides are known for a number of host cells, including, without limitation, bacterial, yeast, mammalian, insect and plant cells.

Techniques for transforming a wide variety of higher plant species are well known and described in the technical and scientific literature. See, e.g. Weising et al., *Ann. Rev. Genet.* 22:421 (1988); and Christou, *Euphytica*, v. 85, n.1-3:13-27, (1995).

DNA constructs of the invention may be introduced into the genome of the desired plant host by a variety of conventional techniques. For example, the DNA construct may be introduced directly into the genomic DNA of the plant cell using techniques such as electroporation and microinjection of plant cell protoplasts, or the DNA constructs can be introduced directly to plant tissue using ballistic methods, such as DNA particle bombardment. Alternatively, the DNA constructs may be combined with suitable T-DNA flanking regions and introduced into a conventional *Agrobacterium tumefaciens* host vector. The virulence functions of the *Agrobacterium tumefaciens* host will direct the insertion of the construct and adjacent marker into the plant cell DNA when the cell is infected by the bacteria (McCormac et al., *Mol. Biotechnol.* 8:199 (1997); Hamilton, *Gene* 200:107 (1997)); Salomon et al. *EMBO J.* 3:141 (1984); Herrera-Estrella et al. *EMBO J.* 2:987 (1983).

Microinjection techniques are known in the art and well described in the scientific and patent literature. The introduction of DNA constructs using polyethylene glycol precipitation is described in Paszkowski et al. *EMBO J.* 3:2717 (1984). Electroporation techniques are described in Fromm et al. *Proc. Natl Acad. Sci. USA* 82:5824 (1985). Ballistic transformation techniques are described in Klein et al. *Nature* 327:773 (1987). *Agrobacterium tumefaciens*-mediated transformation techniques, including disarming and use of binary or co-integrate vectors, are well described in the scientific literature. See, for example Hamilton, *CM., Gene* 200:107 (1997); Müller et al. *Mol. Gen. Genet.* 207:171 (1987); Komari et al. *Plant J.* 10:165 (1996); Venkateswarlu et al. *Biotechnology* 9:1103 (1991) and Gleave, *AP., Plant Mol. Biol.* 20:1203 (1992); Graves and Goldman, *Plant Mol. Biol.* 7:34 (1986) and Gould et al., *Plant Physiology* 95:426 (1991).

Transformed plant cells which are derived by any of the above transformation techniques can be cultured to regenerate a whole plant that possesses the transformed genotype and thus the desired phenotype such as seedlessness. Such regeneration techniques rely on manipulation of certain phytohormones in a tissue culture growth medium, typically relying on a biocide and/or herbicide marker which has been introduced together with the desired nucleotide sequences. Plant regeneration from cultured protoplasts is described in Evans et al., *Protoplasts Isolation and Culture* in "Handbook of Plant Cell Culture," pp. 124-176, MacMillan Publishing Company, New York, 1983; and Binding, *Regeneration of Plants, Plant Protoplasts*, pp. 21-73,

CRC Press, Boca Raton, 1988. Regeneration can also be obtained from plant callus, explants, organs, or parts thereof. Such regeneration techniques are described generally in Klee et al. *Ann. Rev. of Plant Phys.* 38:467 (1987). Regeneration of monocots (rice) is described by Hosoyama et al. (*Biosci. Biotechnol. Biochem.* 58:1500 (1994)) and by Ghosh et al. (*J. Biotechnol.* 32:1 (1994)). The nucleic acids of the invention can be used to confer desired traits on essentially any plant.

Thus, the invention has use over a broad range of plants, including species from the genera *Anacardium*, *Arachis*, *Asparagus*, *Atropa*, *Avena*, *Brassica*, *Citrus*, *Citrullus*, *Capsicum*, *Carthamus*, *Cocos*, *Coffea*, *Cucumis*, *Cucurbita*, *Daucus*, *Elaeis*, *Fragaria*, *Glycine*, *Gossypium*, *Helianthus*, *Heterocallis*, *Hordeum*, *Hyoscyamus*, *Lactuca*, *Linum*, *Lolium*, *Lupinus*, *Lycopersicon*, *Malus*, *Manihot*, *Majorana*, *Medicago*, *Nicotiana*, *Olea*, *Oryza*, *Panicum*, *Pannisetum*, *Persea*, *Phaseolus*, *Pistachia*, *Pisum*, *Pyrus*, *Prunus*, *Raphanus*, *Ricinus*, *Secale*, *Senecio*, *Sinapis*, *Solanum*, *Sorghum*, *Theobromus*, *Trigonella*, *Triticum*, *Vicia*, *Vitis*, *Vigna*, and, *Zea*.

One of skill will recognize that after the expression cassette is stably incorporated in transgenic plants and confirmed to be operable, it can be introduced into other plants by sexual crossing. Any of a number of standard breeding techniques can be used, depending upon the species to be crossed.

The particular sequences of SDFs identified are provided in the attached Tables 1 and 2. One of ordinary skill in the art, having this data, can obtain cloned DNA fragments, synthetic DNA fragments or polypeptides constituting desired sequences by recombinant methodology known in the art or described herein.

EXAMPLES

The invention is illustrated by way of the following examples. The invention is not limited by these examples as the scope of the invention is defined solely by the claims following.

EXAMPLE 1: cDNA PREPARATION

A number of the nucleotide sequences disclosed in Tables 1 and 2 herein as representative of the SDFs of the invention can be obtained by sequencing genomic DNA (gDNA) and/or cDNA from corn plants grown from HYBRID SEED # 35A19, purchased from

Pioneer Hi-Bred International, Inc., Supply Management, P.O. Box 256, Johnston, Iowa 50131-0256.

A number of the nucleotide sequences disclosed in Tables 1 and 2 herein as representative of the SDFs of the invention can also be obtained by sequencing genomic DNA from *Arabidopsis thaliana*, Wassilewskija ecotype or by sequencing cDNA obtained from mRNA from such plants as described below. This is a true breeding strain. Seeds of the plant are available from the Arabidopsis Biological Resource Center at the Ohio State University, under the accession number CS2360. Seeds of this plant were deposited under the terms and conditions of the Budapest Treaty at the American Type Culture Collection, Manassas, VA on August 31, 1999, and were assigned ATCC No. PTA-595.

Other methods for cloning full-length cDNA are described, for example, by Seki et al., *Plant Journal* 15:707-720 (1998) "High-efficiency cloning of Arabidopsis full-length cDNA by biotinylated Cap trapper"; Maruyama et al., *Gene* 138:171 (1994) "Oligo-capping a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides"; and WO 96/34981.

Tissues were, or each organ was, individually pulverized and frozen in liquid nitrogen. Next, the samples were homogenized in the presence of detergents and then centrifuged. The debris and nuclei were removed from the sample and more detergents were added to the sample. The sample was centrifuged and the debris was removed. Then the sample was applied to a 2M sucrose cushion to isolate polysomes. The RNA was isolated by treatment with detergents and proteinase K followed by ethanol precipitation and centrifugation. The polysomal RNA from the different tissues was pooled according to the following mass ratios: 15/15/1 for male inflorescences, female inflorescences and root, respectively. The pooled material was then used for cDNA synthesis by the methods described below.

Starting material for cDNA synthesis for the exemplary corn cDNA clones with sequences presented in Tables 1 and 2 was poly(A)-containing polysomal mRNAs from inflorescences and root tissues of corn plants grown from HYBRID SEED # 35A19. Male inflorescences and female (pre-and post-fertilization) inflorescences were isolated at various stages of development. Selection for poly(A) containing polysomal RNA was done using oligo d(T) cellulose columns, as described by Cox and Goldberg, "Plant Molecular Biology: A Practical Approach", pp. 1-35, Shaw ed., c. 1988 by IRL, Oxford. The quality and the integrity of the polyA+ RNAs were evaluated.

Starting material for cDNA synthesis for the exemplary *Arabidopsis* cDNA clones with sequences presented in Tables 1 and 2 was polysomal RNA isolated from the top-most inflorescence tissues of *Arabidopsis thaliana* Wassilewskija (Ws.) and from roots of *Arabidopsis thaliana* Landsberg erecta (L. er.), also obtained from the Arabidopsis Biological Resource Center. Nine parts inflorescence to every part root was used, as measured by wet mass. Tissue was pulverized and exposed to liquid nitrogen. Next, the sample was homogenized in the presence of detergents and then centrifuged. The debris and nuclei were removed from the sample and more detergents were added to the sample. The sample was centrifuged and the debris was removed and the sample was applied to a 2M sucrose cushion to isolate polysomal RNA. Cox et al., "Plant Molecular Biology: A Practical Approach", pp. 1-35, Shaw ed., c. 1988 by IRL, Oxford. The polysomal RNA was used for cDNA synthesis by the methods described below. Polysomal mRNA was then isolated as described above for corn cDNA. The quality of the RNA was assessed electrophoretically.

Following preparation of the mRNAs from various tissues as described above, selection of mRNA with intact 5' ends and specific attachment of an oligonucleotide tag to the 5' end of such mRNA was performed using either a chemical or enzymatic approach. Both techniques take advantage of the presence of the "cap" structure, which characterizes the 5' end of most intact mRNAs and which comprises a guanosine generally methylated once, at the 7 position.

The chemical modification approach involves the optional elimination of the 2', 3'-cis diol of the 3' terminal ribose, the oxidation of the 2', 3'-cis diol of the ribose linked to the cap of the 5' ends of the mRNAs into a dialdehyde, and the coupling of the such obtained dialdehyde to a derivatized oligonucleotide tag. Further detail regarding the chemical approaches for obtaining mRNAs having intact 5' ends are disclosed in International Application No.

WO96/34981 published November 7, 1996.

The enzymatic approach for ligating the oligonucleotide tag to the intact 5' ends of mRNAs involves the removal of the phosphate groups present on the 5' ends of uncapped incomplete mRNAs, the subsequent decapping of mRNAs having intact 5' ends and the ligation of the phosphate present at the 5' end of the decapped mRNA to an oligonucleotide tag. Further detail regarding the enzymatic approaches for obtaining mRNAs having intact 5' ends are disclosed in Dumas Milne Edwards J.B. (Doctoral Thesis of Paris VI University, Le clonage des ADNc complets: difficultés et perspectives nouvelles. Apports pour l'étude de la régulation de

l'expression de la tryptophane hydroxylase de rat, 20 Dec. 1993), EP0 625572 and Kato *et al.*, *Gene* 150:243-250 (1994).

In both the chemical and the enzymatic approach, the oligonucleotide tag has a restriction enzyme site (e.g. an EcoRI site) therein to facilitate later cloning procedures.

5 Following attachment of the oligonucleotide tag to the mRNA, the integrity of the mRNA is examined by performing a Northern blot using a probe complementary to the oligonucleotide tag.

10 For the mRNAs joined to oligonucleotide tags using either the chemical or the enzymatic method, first strand cDNA synthesis is performed using an oligo-dT primer with reverse transcriptase. This oligo-dT primer can contain an internal tag of at least 4 nucleotides, which can be different from one mRNA preparation to another. Methylated dCTP is used for cDNA first strand synthesis to protect the internal EcoRI sites from digestion during subsequent steps. The first strand cDNA is precipitated using isopropanol after removal of RNA by alkaline hydrolysis to eliminate residual primers.

15 Second strand cDNA synthesis is conducted using a DNA polymerase, such as Klenow fragment and a primer corresponding to the 5' end of the ligated oligonucleotide. The primer is typically 20-25 bases in length. Methylated dCTP is used for second strand synthesis in order to protect internal EcoRI sites in the cDNA from digestion during the cloning process.

20 Following second strand synthesis, the full-length cDNAs are cloned into a phagemid vector, such as pBlueScriptTM (Stratagene). The ends of the full-length cDNAs are blunted with T4 DNA polymerase (Biolabs) and the cDNA is digested with EcoRI. Since methylated dCTP is used during cDNA synthesis, the EcoRI site present in the tag is the only hemi-methylated site; hence the only site susceptible to EcoRI digestion. In some instances, to facilitate subcloning, an Hind III adapter is added to the 3' end of full-length cDNAs.

25 The full-length cDNAs are then size fractionated using either exclusion chromatography (AcA, Biosepra) or electrophoretic separation which yields 3 to 6 different fractions. The full-length cDNAs are then directionally cloned either into pBlueScriptTM using either the EcoRI and SmaI restriction sites or, when the Hind III adapter is present in the full-length cDNAs, the EcoRI and Hind III restriction sites. The ligation mixture is transformed, preferably by
30 electroporation, into bacteria, which are then propagated under appropriate antibiotic selection.

Clones containing the oligonucleotide tag attached to full-length cDNAs are selected as follows.

The plasmid cDNA libraries made as described above are purified (e.g. by a column available from Qiagen). A positive selection of the tagged clones is performed as follows. Briefly, in this selection procedure, the plasmid DNA is converted to single stranded DNA using phage F1 gene II endonuclease in combination with an exonuclease (Chang et al., *Gene* 127:95 (1993)) such as exonuclease III or T7 gene 6 exonuclease. The resulting single stranded DNA is then purified using paramagnetic beads as described by Fry et al., *Biotechniques* 13: 124 (1992). Here the single stranded DNA is hybridized with a biotinylated oligonucleotide having a sequence corresponding to the 3' end of the oligonucleotide tag. Preferably, the primer has a length of 20-25 bases. Clones including a sequence complementary to the biotinylated oligonucleotide are selected by incubation with streptavidin coated magnetic beads followed by magnetic capture. After capture of the positive clones, the plasmid DNA is released from the magnetic beads and converted into double stranded DNA using a DNA polymerase such as ThermoSequenase™ (obtained from Amersham Pharmacia Biotech). Alternatively, protocols such as the Gene Trapper™ kit (Gibco BRL) can be used. The double stranded DNA is then transformed, preferably by electroporation, into bacteria. The percentage of positive clones having the 5' tag oligonucleotide is typically estimated to be between 90 and 98% from dot blot analysis.

Following transformation, the libraries are ordered in microtiter plates and sequenced. The *Arabidopsis* library was deposited at the American Type Culture Collection on January 7, 2000 as "*E-coli* liba 010600" under the accession number PTA-1161.

EXAMPLE 2: SOUTHERN HYBRIDIZATIONS

The SDFs of the invention can be used in Southern hybridizations as described above. The following describes extraction of DNA from nuclei of plant cells, digestion of the nuclear DNA and separation by length, transfer of the separated fragments to membranes, preparation of probes for hybridization, hybridization and detection of the hybridized probe.

The procedures described herein can be used to isolate related polynucleotides or for diagnostic purposes. Moderate stringency hybridization conditions, as defined above, are described in the present example. These conditions result in detection of hybridization between sequences having at least 70% sequence identity. As described above, the hybridization and wash conditions can be changed to reflect the desired percentatge of sequence identity between probe and target sequences that can be detected.

1031

In the following procedure, a probe for hybridization is produced from two PCR reactions using two primers from genomic sequence of *Arabidopsis thaliana*. As described above, the particular template for generating the probe can be any desired template.

The first PCR product is assessed to validate the size of the primer to assure it is of the expected size. Then the product of the first PCR is used as a template, with the same pair of primers used in the first PCR, in a second PCR that produces a labeled product used as the probe.

Fragments detected by hybridization, or other bands of interest, can be isolated from gels used to separate genomic DNA fragments by known methods for further purification and/or characterization.

Buffers for nuclear DNA extraction

1. 10X HB

	1000 ml	
40 mM spermidine	10.2 g	Spermine (Sigma S-2876) and spermidine (Sigma S-2501)
10 mM spermine	3.5 g	Stabilize chromatin and the nuclear membrane
0.1 M EDTA (disodium)	37.2 g	EDTA inhibits nuclease
0.1 M Tris	12.1 g	Buffer
0.8 M KCl	59.6 g	Adjusts ionic strength for stability of nuclei

Adjust pH to 9.5 with 10 N NaOH. It appears that there is a nuclease present in leaves. Use of pH 9.5 appears to inactivate this nuclease.

2. 2 M sucrose (684 g per 1000 ml)

1032

Heat about half the final volume of water to about 50°C. Add the sucrose slowly then bring the mixture to close to final volume; stir constantly until it has dissolved. Bring the solution to volume.

3. Sarkosyl solution (lyses nuclear membranes)

5		<u>1000 ml</u>
	N-lauroyl sarcosine (Sarkosyl)	20.0 g
	0.1 M Tris	12.1 g
	0.04 M EDTA (Disodium)	14.9 g

Adjust the pH to 9.5 after all the components are dissolved and bring up to the proper volume.

4. 20% Triton X-100
80 ml Triton X-100
320 ml 1xHB (w/o β-ME and PMSF)
Prepare in advance; Triton takes some time to dissolve

A. Procedure

1. Prepare 1X "H" buffer (keep ice-cold during use)

1000 ml

10X HB	100 ml
2 M sucrose	250 ml a non-ionic osmoticum
Water	634 ml

Added just before use:

100 mM PMSF*	10 ml a protease inhibitor; protects nuclear membrane proteins
β-mercaptoethanol	1 ml inactivates nuclease by reducing disulfide bonds

*100 mM PMSF

(phenyl methyl sulfonyl fluoride, Sigma P-7626)

(add 0.0875 g to 5 ml 100% ethanol)

2. Homogenize the tissue in a blender (use 300-400 ml of 1xHB per blender). Be sure
5 that you use 5-10 ml of HB buffer per gram of tissue. Blenders generate heat so be
sure to keep the homogenate cold. It is necessary to put the blenders in ice
periodically.
3. Add the 20% Triton X-100 (25 ml per liter of homogenate) and gently stir on ice for
20 min. This lyses plastid, but not nuclear, membranes.
4. Filter the tissue suspension through several nylon filters into an ice-cold beaker. The
first filtration is through a 250-micron membrane; the second is through an 85-micron
membrane; the third is through a 50-micron membrane; and the fourth is through a
20-micron membrane. Use a large funnel to hold the filters. Filtration can be sped up
by gently squeezing the liquid through the filters.
5. Centrifuge the filtrate at 1200 x g for 20 min. at 4°C to pellet the nuclei.
6. Discard the dark green supernatant. The pellet will have several layers to it. One is
starch; it is white and gritty. The nuclei are gray and soft. In the early steps, there
may be a dark green and somewhat viscous layer of chloroplasts.

Wash the pellets in about 25 ml cold H buffer (with Triton X-100) and resuspend by
20 swirling gently and pipetting. After the pellets are resuspended.

Pellet the nuclei again at 1200 - 1300 x g. Discard the supernatant.

Repeat the wash 3-4 times until the supernatant has changed from a dark green to a
pale green. This usually happens after 3 or 4 resuspensions. At this point, the pellet

1034

is typically grayish white and very slippery. The Triton X-100 in these repeated steps helps to destroy the chloroplasts and mitochondria that contaminate the prep.

Resuspend the nuclei for a final time in a total of 15 ml of H buffer and transfer the suspension to a sterile 125 ml Erlenmeyer flask.

- 5 7. Add 15 ml, dropwise, cold 2% Sarkosyl, 0.1 M Tris, 0.04 M EDTA solution (pH 9.5) while swirling gently. This lyses the nuclei. The solution will become very viscous.
8. Add 30 grams of CsCl and gently swirl at room temperature until the CsCl is in solution. The mixture will be gray, white and viscous.
9. Centrifuge the solution at 11,400 x g at 4°C for at least 30 min. The longer this spin is, the firmer the protein pellicle.
10. The result is typically a clear green supernatant over a white pellet, and (perhaps) under a protein pellicle. Carefully remove the solution under the protein pellicle and above the pellet. Determine the density of the solution by weighing 1 ml of solution and add CsCl if necessary to bring to 1.57 g/ml. The solution contains dissolved solids (sucrose etc) and the refractive index alone will not be an accurate guide to CsCl concentration.
11. Add 20 µl of 10 mg/ml EtBr per ml of solution.
12. Centrifuge at 184,000 x g for 16 to 20 hours in a fixed-angle rotor.
- 20 13. Remove the dark red supernatant that is at the top of the tube with a plastic transfer pipette and discard. Carefully remove the DNA band with another transfer pipette. The DNA band is usually visible in room light; otherwise, use a long wave UV light to locate the band.
- 25 14. Extract the ethidium bromide with isopropanol saturated with water and salt. Once the solution is clear, extract at least two more times to ensure that all of the EtBr is

1035

gone. Be very gentle, as it is very easy to shear the DNA at this step. This extraction may take a while because the DNA solution tends to be very viscous. If the solution is too viscous, dilute it with TE.

15. Dialyze the DNA for at least two days against several changes (at least three times) of TE (10 mM Tris, 1mM EDTA, pH 8) to remove the cesium chloride.
16. Remove the dialyzed DNA from the tubing. If the dialyzed DNA solution contains a lot of debris, centrifuge the DNA solution at least at 2500 x g for 10 min. and carefully transfer the clear supernatant to a new tube. Read the A260 concentration of the DNA.
17. Assess the quality of the DNA by agarose gel electrophoresis (1% agarose gel) of the DNA. Load 50 ng and 100 ng (based on the OD reading) and compare it with known and good quality DNA. Undigested lambda DNA and a lambda-HindIII-digested DNA are good molecular weight makers.

Protocol for Digestion of Genomic DNA

Protocol:

1. The relative amounts of DNA for different crop plants that provide approximately a balanced number of genome equivalent is given in Table 3. Note that due to the size of the wheat genome, wheat DNA will be underrepresented. Lambda DNA provides a useful control for complete digestion.
2. Precipitate the DNA by adding 3 volumes of 100% ethanol. Incubate at -20°C for at least two hours. Yeast DNA can be purchased and made up at the necessary concentration, therefore no precipitation is necessary for yeast DNA.
3. Centrifuge the solution at 11,400 x g for 20 min. Decant the ethanol carefully (be careful not to disturb the pellet). Be sure that the residual ethanol is completely removed either by vacuum desiccation or by carefully wiping the sides of the tubes with a clean tissue.

1036

4. Resuspend the pellet in an appropriate volume of water. Be sure the pellet is fully resuspended before proceeding to the next step. This may take about 30 min.
5. Add the appropriate volume of 10X reaction buffer provided by the manufacturer of the restriction enzyme to the resuspended DNA followed by the appropriate volume of enzymes. Be sure to mix it properly by slowly swirling the tubes.
6. Set-up the lambda digestion-control for each DNA that you are digesting.
7. Incubate both the experimental and lambda digests overnight at 37°C. Spin down condensation in a microfuge before proceeding.
8. After digestion, add 2 µl of loading dye (typically 0.25% bromophenol blue, 0.25% xylene cyanol in 15% Ficoll or 30% glycerol) to the lambda-control digests and load in 1% TPE-agarose gel (TPE is 90 mM Tris-phosphate, 2 mM EDTA, pH 8). If the lambda DNA in the lambda control digests are completely digested, proceed with the precipitation of the genomic DNA in the digests.
9. Precipitate the digested DNA by adding 3 volumes of 100% ethanol and incubating in -20°C for at least 2 hours (preferably overnight).

EXCEPTION: *Arabidopsis* and yeast DNA are digested in an appropriate volume; they don't have to be precipitated.

10. Resuspend the DNA in an appropriate volume of TE (e.g., 22 µl x 50 blots = 1100 µl) and an appropriate volume of 10X loading dye (e.g., 2.4 µl x 50 blots = 120 µl). Be careful in pipetting the loading dye - it is viscous. Be sure you are pipetting the correct volume.

Table 3

Some guide points in digesting genomic DNA.

			Genome	Amount

1037

Species	Genome Size	Size Relative to Arabidopsis	Equivalent to 2 μ g Arabidopsis DNA	of DNA per blot
Arabidopsis	120 Mb	1X	1X	2 μ g
Brassica	1,100 Mb	9.2X	0.54X	10 μ g
Corn	2,800 Mb	23.3X	0.43X	20 μ g
Cotton	2,300 Mb	19.2X	0.52X	20 μ g
Oat	11,300 Mb	94X	0.11X	20 μ g
Rice	400 Mb	3.3X	0.75X	5 μ g
Soybean	1,100 Mb	9.2X	0.54X	10 μ g
Sugarbeet	758 Mb	6.3X	0.8X	10 μ g
Sweetclover	1,100 Mb	9.2X	0.54X	10 μ g
Wheat	16,000 Mb	133X	0.08X	20 μ g
Yeast	15 Mb	0.12X	1X	0.25 μ g

Protocol for Southern Blot Analysis

The digested DNA samples are electrophoresed in 1% agarose gels in 1x TPE buffer. Low voltage; overnight separations are preferred. The gels are stained with EtBr and photographed.

1. For blotting the gels, first incubate the gel in 0.25 N HCl (with gentle shaking) for about 15 min.
2. Then briefly rinse with water. The DNA is denatured by 2 incubations. Incubate (with shaking) in 0.5 M NaOH in 1.5 M NaCl for 15 min.
3. The gel is then briefly rinsed in water and neutralized by incubating twice (with shaking) in 1.5 M Tris pH 7.5 in 1.5 M NaCl for 15 min.

1038

4. A nylon membrane is prepared by soaking it in water for at least 5 min, then in 6X SSC for at least 15 min. before use. (20x SSC is 175.3 g NaCl, 88.2 g sodium citrate per liter, adjusted to pH 7.0.)
5. The nylon membrane is placed on top of the gel and all bubbles in between are removed. The DNA is blotted from the gel to the membrane using an absorbent medium, such as paper toweling and 6x SCC buffer. After the transfer, the membrane may be lightly brushed with a gloved hand to remove any agarose sticking to the surface.
6. The DNA is then fixed to the membrane by UV crosslinking and baking at 80°C. The membrane is stored at 4°C until use.

B. Protocol for PCR Amplification of Genomic Fragments in Arabidopsis

Amplification procedures:

1. Mix the following in a 0.20 ml PCR tube or 96-well PCR plate:

Volume	Stock	Final Amount or Conc.
0.5 µl	~ 10 ng/µl genomic DNA ¹	5 ng
2.5 µl	10X PCR buffer	20 mM Tris, 50 mM KCl
0.75 µl	50 mM MgCl ₂	1.5 mM
1 µl	10 pmol/µl Primer 1 (Forward)	10 pmol
1 µl	10 pmol/µl Primer 2 (Reverse)	10 pmol
0.5 µl	5 mM dNTPs	0.1 mM

¹ Arabidopsis DNA is used in the present experiment, but the procedure is a general one.

1039

0.1 µl	5 units/µl Platinum Taq™ (Life Technologies, Gaithersburg, MD) DNA Polymerase	1 units
(to 25 µl)	Water	

2. The template DNA is amplified using a Perkin Elmer 9700 PCR machine:

1) 94°C for 10 min. followed by

<u>2)</u> 5 cycles:	<u>3)</u> 5 cycles:	<u>4)</u> 25 cycles:
94 °C - 30 sec 62 °C - 30 sec 72 °C - 3 min	94 °C - 30 sec 58 °C - 30 sec 72 °C - 3 min	94 °C - 30 sec 53 °C - 30 sec 72 °C - 3 min

5) 72°C for 7 min. Then the reactions are stopped by chilling to 4°C.

The procedure can be adapted to a multi-well format if necessary.

5 Quantification and Dilution of PCR Products:

1. The product of the PCR is analyzed by electrophoresis in a 1% agarose gel. A linearized plasmid DNA can be used as a quantification standard (usually at 50, 100, 200, and 400 ng). These will be used as references to approximate the amount of PCR products. HindIII-digested Lambda DNA is useful as a molecular weight marker. The gel can be run fairly quickly; e.g., at 100 volts. The standard gel is examined to determine that the size of the PCR products is consistent with the expected size and if there are significant extra bands or smeary products in the PCR reactions.

2. The amounts of PCR products can be estimated on the basis of the plasmid standard.

3. For the small number of reactions that produce extraneous bands, a small amount of DNA from bands with the correct size can be isolated by dipping a sterile 10- μ l tip into the band while viewing through a UV Transilluminator. The small amount of agarose gel (with the DNA fragment) is used in the labeling reaction.

5 C. Protocol for PCR-DIG-Labeling of DNA

Solutions:

Reagents in PCR reactions (diluted PCR products, 10X PCR Buffer, 50 mM MgCl₂, 5 U/ μ l Platinum Taq Polymerase, and the primers)

10X dNTP + DIG-11-dUTP [1:5]: (2 mM dATP, 2 mM dCTP, 2 mM dGTP, 1.65 mM dTTP, 0.35 mM DIG-11-dUTP)

10X dNTP + DIG-11-dUTP [1:10]: (2 mM dATP, 2 mM dCTP, 2 mM dGTP, 1.81 mM dTTP, 0.19 mM DIG-11-dUTP)

10X dNTP + DIG-11-dUTP [1:15]: (2 mM dATP, 2 mM dCTP, 2 mM dGTP, 1.875 mM dTTP, 0.125 mM DIG-11-dUTP)

TE buffer (10 mM Tris, 1 mM EDTA, pH 8)

Maleate buffer: In 700 ml of deionized distilled water, dissolve 11.61 g maleic acid and 8.77 g NaCl. Add NaOH to adjust the pH to 7.5. Bring the volume to 1 L. Stir for 15 min. and sterilize.

10% blocking solution: In 80 ml deionized distilled water, dissolve 1.16g maleic acid. Next, add NaOH to adjust the pH to 7.5. Add 10 g of the blocking reagent powder (Boehringer Mannheim, Indianapolis, IN, Cat. no. 1096176). Heat to 60°C while stirring to dissolve the powder. Adjust the volume to 100 ml with water. Stir and sterilize.

1% blocking solution: Dilute the 10% stock to 1% using the maleate buffer.

[illegible]

Procedure:

1. PCR reactions are performed in 25 µl volumes containing:

PCR buffer	1X
MgCl ₂	1.5 mM
10X dNTP + DIG-11-dUTP	1X (please see the note below)
Platinum Taq™ Polymerase	1 unit
10 pg probe DNA	
10 pmol primer 1	

Note:Use for:

10X dNTP + DIG-11-dUTP (1:5)	< 1 kb
10X dNTP + DIG-11-dUTP (1:10)	1 kb to 1.8 kb
10X dNTP + DIG-11-dUTP (1:15)	> 1.8 kb

2. The PCR reaction uses the following amplification cycles:

- 1) 94°C for 10 min.

<u>2)</u> 5 cycles:	<u>3)</u> 5 cycles:	<u>4)</u> 25 cycles:
95°C - 30 sec 61°C - 1 min 73°C - 5 min	95°C - 30 sec 59°C - 1 min 75°C - 5 min	95°C - 30 sec 51°C - 1 min 73°C - 5 min

- 5) 72°C for 8 min. The reactions are terminated by chilling to 4°C (hold).

3. The products are analyzed by electrophoresis- in a 1% agarose gel, comparing to an aliquot of the unlabelled probe starting material.
4. The amount of DIG-labeled probe is determined as follows:

1043

Make serial dilutions of the diluted control DNA in dilution buffer (TE: 10 mM Tris and 1 mM EDTA, pH 8) as shown in the following table:

DIG-labeled control DNA starting conc.	Stepwise Dilution	Final Conc. (Dilution Name)
5 ng/ μ l	1 μ l in 49 μ l TE	100 pg/ μ l (A)
100 pg/ μ l (A)	25 μ l in 25 μ l TE	50 pg/ μ l (B)
50 pg/ μ l (B)	25 μ l in 25 μ l TE	25 pg/ μ l (C)
25 pg/ μ l (C)	20 μ l in 30 μ l TE	10 pg/ μ l (D)

- a. Serial dilutions of a DIG-labeled standard DNA ranging from 100 pg to 10 pg are spotted onto a positively charged nylon membrane, marking the membrane lightly with a pencil to identify each dilution.
- b. Serial dilutions (e.g., 1:50, 1:2500, 1:10,000) of the newly labeled DNA probe are spotted.
- c. The membrane is fixed by UV crosslinking.
- d. The membrane is wetted with a small amount of maleate buffer and then incubated in 1% blocking solution for 15 min at room temp.
- e. The labeled DNA is then detected using alkaline phosphatase conjugated anti-DIG antibody (Boehringer Mannheim, Indianapolis, IN, cat. no. 1093274) and an NBT substrate according to the manufacture's instruction.
- f. Spot intensities of the control and experimental dilutions are then compared to estimate the concentration of the PCR-DIG-labeled probe.

D. Prehybridization and Hybridization of Southern BlotsSolutions:

100% Formamide purchased from Gibco

20X SSC (1X = 0.15 M NaCl, 0.015 M Na₃citrate)

per L: 175 g NaCl
87.5 g Na₃citrate·2H₂O

20% Sarkosyl (N-lauroyl-sarcosine)

20% SDS (sodium dodecyl sulphate)

10% Blocking Reagent: In 80 ml deionized distilled water, dissolve 1.16 g maleic acid. Next, add NaOH to adjust the pH to 7.5. Add 10 g of the blocking reagent powder. Heat to 60°C while stirring to dissolve the powder. Adjust the volume to 100 ml with water. Stir and sterilize.

Prehybridization Mix:

Final Concentration	Components	Volume (per 100 ml)	Stock
50%	Formamide	50 ml	100%
5X	SSC	25 ml	20X
0.1%	Sarkosyl	0.5 ml	20%
0.02%	SDS	0.1 ml	20%
2%	Blocking Reagent	20 ml	10%
	Water	4.4 ml	

General Procedures:

- Place the blot in a heat-sealable plastic bag and add an appropriate volume of prehybridization solution (30 ml/100cm²) at room temperature. Seal the bag with a heat sealer, avoiding bubbles as much as possible. Lay down the bags in a large plastic tray (one tray can accommodate at least 4–5 bags). Ensure that the bags are

1045

lying flat in the tray so that the prehybridization solution is evenly distributed throughout the bag. Incubate the blot for at least 2 hours with gentle agitation using a waver shaker.

2. Denature DIG-labeled DNA probe by incubating for 10 min. at 98°C using the PCR machine and immediately cool it to 4°C.

3. Add probe to prehybridization solution (25 ng/ml; 30 ml = 750 ng total probe) and mix well but avoid foaming. Bubbles may lead to background.

4. Pour off the prehybridization solution from the hybridization bags and add new prehybridization and probe solution mixture to the bags containing the membrane.

5. Incubate with gentle agitation for at least 16 hours.

6. Proceed to medium stringency post-hybridization wash:

Three times for 20 min. each with gentle agitation using 1X SSC, 1% SDS at 60°C.

All wash solutions must be prewarmed to 60°C. Use about 100 ml of wash solution per membrane.

7. To avoid background keep the membranes fully submerged to avoid drying in spots; agitate sufficiently to avoid having membranes stick to one another.

7. After the wash, proceed to immunological detection and CSPD development.

E. Procedure for Immunological Detection with CSPD

Solutions:

Buffer 1: Maleic acid buffer (0.1 M maleic acid, 0.15 M NaCl; adjusted to pH 7.5 with NaOH)

1046

Washing buffer: Maleic acid buffer with 0.3% (v/v) Tween 20.

Blocking stock solution 10% blocking reagent in buffer 1. Dissolve (10X concentration): blocking reagent powder (Boehringer Mannheim, Indianapolis, IN, cat. no. 1096176) by constantly stirring on a 65°C heating block or heat in a microwave, autoclave and store at 4°C.

Buffer 2

(1X blocking solution): Dilute the stock solution 1:10 in Buffer 1.

Detection buffer: 0.1 M Tris, 0.1 M NaCl, pH 9.5

Procedure:

1. After the post-hybridization wash the blots are briefly rinsed (1-5 min.) in the maleate washing buffer with gentle shaking.
2. Then the membranes are incubated for 30 min. in Buffer 2 with gentle shaking.
3. Anti-DIG-AP conjugate (Boehringer Mannheim, Indianapolis, IN, cat. no. 1093274) at 75 mU/ml (1:10,000) in Buffer 2 is used for detection. 75 ml of solution can be used for 3 blots.
4. The membrane is incubated for 30 min. in the antibody solution with gentle shaking.
5. The membrane are washed twice in washing buffer with gentle shaking. About 250 mls is used per wash for 3 blots.
6. The blots are equilibrated for 2-5 min in 60 ml detection buffer.
7. Dilute CSPD (1:200) in detection buffer. (This can be prepared ahead of time and stored in the dark at 4°C).

The following steps must be done individually. Bags (one for detection and one for exposure) are generally cut and ready before doing the following steps.

8. The blot is carefully removed from the detection buffer and excess liquid removed without drying the membrane. The blot is immediately placed in a bag and 1.5 ml of CSPD solution is added. The CSPD solution can be spread over the membrane. Bubbles present at the edge and on the surface of the blot are typically removed by gentle rubbing. The membrane is incubated for 5 min. in CSPD solution.
9. Excess liquid is removed and the membrane is blotted briefly (DNA side up) on Whatman 3MM paper. Do not let the membrane dry completely.
10. Seal the damp membrane in a hybridization bag and incubate for 10 min at 37°C to enhance the luminescent reaction.
11. Expose for 2 hours at room temperature to X-ray film. Multiple exposures can be taken. Luminescence continues for at least 24 hours and signal intensity increases during the first hours.

Example 3: Transformation of Carrot Cells

Transformation of plant cells can be accomplished by a number of methods, as described above. Similarly, a number of plant genera can be regenerated from tissue culture following transformation. Transformation and regeneration of carrot cells as described herein is illustrative.

Single cell suspension cultures of carrot (*Daucus carota*) cells are established from hypocotyls of cultivar Early Nantes in B₅ growth medium (O.L. Gamborg et al., *Plant Physiol.* **45**:372 (1970)) plus 2,4-D and 15 mM CaCl₂ (B₅-44 medium) by methods known in the art. The suspension cultures are subcultured by adding 10 ml of the suspension culture to 40 ml of B₅-44 medium in 250 ml flasks every 7 days and are maintained in a shaker at 150 rpm at 27 °C in the dark.

The suspension culture cells are transformed with exogenous DNA as described by Z. Chen et al. *Plant Mol. Bio.* **36**:163 (1998). Briefly, 4-days post-subculture cells are incubated with cell wall digestion solution containing 0.4 M sorbitol, 2% driselase, 5mM MES (2-[N-

Morpholino] ethanesulfonic acid) pH 5.0 for 5 hours. The digested cells are pelleted gently at 60 xg for 5 min. and washed twice in W5 solution containing 154 mM NaCl, 5 mM KCl, 125 mM CaCl₂ and 5mM glucose, pH 6.0. The protoplasts are suspended in MC solution containing 5 mM MES, 20 mM CaCl₂, 0.5 M mannitol, pH 5.7 and the protoplast density is adjusted to about 4×10^6 protoplasts per ml.

15-60 µg of plasmid DNA is mixed with 0.9 ml of protoplasts. The resulting suspension is mixed with 40% polyethylene glycol (MW 8000, PEG 8000), by gentle inversion a few times at room temperature for 5 to 25 min. Protoplast culture medium known in the art is added into the PEG-DNA-protoplast mixture. Protoplasts are incubated in the culture medium for 24 hour to 5 days and cell extracts can be used for assay of transient expression of the introduced gene. Alternatively, transformed cells can be used to produce transgenic callus, which in turn can be used to produce transgenic plants, by methods known in the art. See, for example, Nomura and Komamine, *Plt. Phys.* 79:988-991 (1985), *Identification and Isolation of Single Cells that Produce Somatic Embryos in Carrot Suspension Cultures*.

An additional deposit of an *E. coli* Library, *E. coli*LibA021800, was made at the American Type Culture Collection in Manassas, Virginia, USA on February 22, 2000 to meet the requirements of Budapest Treaty for the international recognition of the deposit of microorganisms.

The invention being thus described, it will be apparent to one of ordinary skill in the art that various modifications of the materials and methods for practicing the invention can be made. Such modifications are to be considered within the scope of the invention as defined by the following claims.

Each of the references from the patent and periodical literature cited herein is hereby expressly incorporated in its entirety by such citation.

CLAIMS

CLAIMS

What is claimed is:

1. An isolated nucleic acid molecule comprising a nucleic acid having a nucleotide sequence which encodes an amino acid sequence exhibiting at least 40% sequence identity to an amino acid sequence encoded by

(a) a nucleotide sequence described in Tables 1 and/or 2 or a fragment thereof;

or

(b) a complement of a nucleotide sequence shown in Tables 1 and/or 2 or a fragment thereof.

2. An isolated nucleic acid molecule comprising a nucleic acid having a nucleotide sequence which exhibits at least 65% sequence identity to

(a) a nucleotide sequence shown in Tables 1 and/or 2 or a fragment thereof; or

(b) a complement of a nucleotide sequence described in Tables 1 and/or 2 or a fragment thereof.

3. An isolated nucleic acid molecule comprising a nucleic acid having a nucleotide sequence which exhibits at least 65% sequence identity to a gene comprising

(a) a nucleotide sequence shown in Tables 1 and/or 2 or a fragment thereof; or

(b) a complement of a nucleotide sequence described in Tables 1 and/or 2 or a fragment thereof.

4. An isolated nucleic acid molecule which is the reverse of the isolated nucleotide sequence according to claim 1, such that the reverse nucleotide sequence has a sequence order which is the reverse of the sequence order of said isolated nucleotide sequence according to claim 1.

5. An isolated nucleic acid molecule comprising a nucleic acid capable of hybridizing to a nucleic acid having a sequence selected from the group consisting of:

(a) a nucleotide sequence which is shown in Tables 1 and/or 2; and

(b) a nucleotide sequence which is complementary to a nucleotide sequence shown in Tables 1 and/or 2;

under conditions that permit formation of a nucleic acid duplex at a temperature from about 40°C and 48°C below the melting temperature of the nucleic acid duplex.

6. The nucleic acid molecule according to claim 1, wherein said nucleic acid comprises an open reading frame.

1050

7. The isolated nucleic acid molecule of claim 1, wherein said nucleic acid is capable of functioning as a promoter, a 3' end termination sequence, an untranslated region (UTR), or as a regulatory sequence.

8. The isolated nucleic acid molecule of claim 7, wherein said nucleic acid is a promoter and comprises a sequence selected from the group consisting of a TATA box sequence, a CAAT box sequence, a motif of GCAATCG or any transcription-factor binding sequence, and any combination thereof.

9. The isolated nucleic acid molecule of claim 7, wherein the nucleic acid sequence is a regulatory sequence which is capable of promoting seed-specific expression, embryo-specific expression, ovule-specific expression, tapetum-specific expression or root-specific expression of a sequence or any combination thereof.

10. A vector construct comprising a nucleic acid molecule according to claim 1, wherein said nucleic acid molecule is heterologous to any element in said vector construct.

11. A vector construct comprising:

(a) a first nucleic acid having a regulatory sequence capable of causing transcription and/or translation; and

(b) a second nucleic acid having the sequence of the isolated nucleic acid molecule according to claim 1;

wherein said first and second nucleic acids are operably linked and

wherein said second nucleic acid is heterologous to any element in said vector construct.

12. The vector construct according to claim 11, wherein said first nucleic acid is native to said second nucleic acid.

13. The vector construct according to claim 11, wherein said first nucleic acid is heterologous to said second nucleic acid.

14. A vector construct comprising:

(c) a first nucleic acid having the sequence of the isolated nucleic acid molecule according to claim 7; and

(d) a second nucleic acid;

wherein said first and second nucleic acids are operably linked and

wherein said first nucleic acid is heterologous to any element in said vector construct.

15. The vector construct according to claim 14, wherein said first nucleic acid is native to said second nucleic acid.

1051

16. The vector construct according to claim 14, wherein said first nucleic acid is heterologous to said second nucleic acid.

17. A host cell comprising an isolated nucleic acid molecule according to claim 1, wherein said nucleic acid molecule is flanked by exogenous sequence.

18. A host cell comprising a vector construct of claim 10.

19. A host cell comprising a vector construct of claim 11.

20. A host cell comprising a vector construct of claim 12.

21. A host cell comprising a vector construct of claim 13.

22. A host cell comprising a vector construct of claim 14.

23. A host cell comprising a vector construct of claim 15.

24. A host cell comprising a vector construct of claim 16.

25. An isolated polypeptide comprising an amino acid sequence

(a) exhibiting at least 40% sequence identity of an amino acid sequence encoded by a sequence shown in Tables 1 and/or 2 or a fragment thereof; and

(b) capable of exhibiting at least one of the biological activities of the polypeptide encoded by said nucleotide sequence shown in Tables 1 and/or 2 or a fragment thereof.

26. The isolated polypeptide of claim 25, wherein said amino acid sequence exhibits at least 75% sequence identity to an amino acid sequence encoded by a sequence shown in Tables 1 and/or 2 or a fragment thereof.

27. The isolated polypeptide of claim 25, wherein said amino acid sequence exhibits at least 85% sequence identity to an amino acid sequence encoded by a sequence shown in Tables 1 and/or 2 or a fragment thereof.

28. The isolated polypeptide of claim 25, wherein said amino acid sequence exhibits at least 90% sequence identity to an amino acid sequence encoded by a sequence shown in Tables 1 and/or 2 or a fragment thereof.

29. An antibody capable of binding the isolated polypeptide of claim 25.

30. A method of introducing an isolated nucleic acid into a host cell comprising:

(a) providing an isolated nucleic acid molecule according to claim 1; and

(b) contacting said isolated nucleic with said host cell under conditions that permit insertion of said nucleic acid into said host cell.

31. A method of transforming a host cell which comprises contacting a host cell with a vector construct according to claim 10.

1052

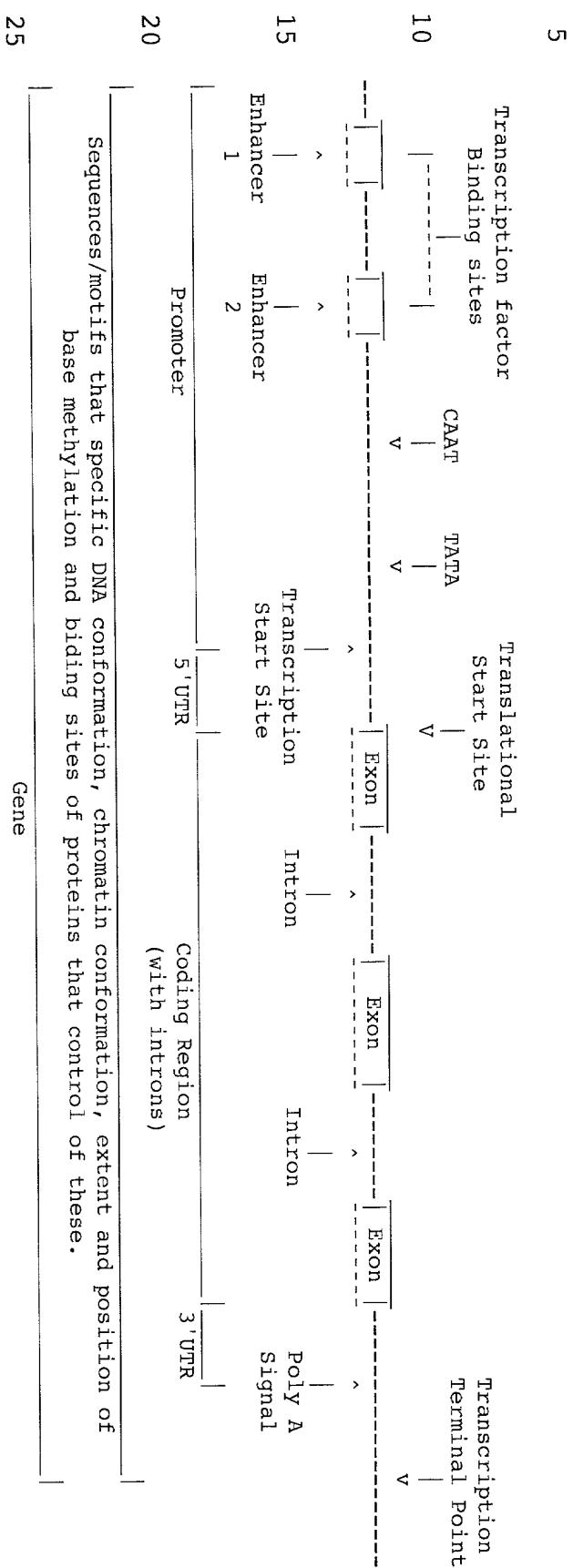
32. A method of transforming a host cell which comprises contacting a host cell with a vector construct according to claim 11.
33. A method of transforming a host cell which comprises contacting a host cell with a vector construct according to claim 12.
34. A method of transforming a host cell which comprises contacting a host cell with a vector construct according to claim 13.
35. A method of transforming a host cell which comprises contacting a host cell with a vector construct according to claim 14.
36. A method of transforming a host cell which comprises contacting a host cell with a vector construct according to claim 15.
37. A method of transforming a host cell which comprises contacting a host cell with a vector construct according to claim 16.
38. A method of modulating transcription and/or translation of a nucleic acid in a host cell comprising:
 - (a) providing the host cell of claim 17; and
 - (b) culturing said host cell under conditions that permit transcription or translation.
39. A method for detecting a nucleic acid in a sample which comprises:
 - (a) providing an isolated nucleic acid molecule according to claim 1;
 - (b) contacting said isolated nucleic acid molecule with a sample under conditions which permit a comparison of the sequence of said isolated nucleic acid molecule with the sequence of DNA in said sample; and
 - (c) analyzing the result of said comparison.
40. The method according to claim 39, wherein said isolated nucleic acid molecule and said sample are contacted under conditions which permit the formation of a duplex between complementary nucleic acid sequences.
41. A plant or cell of a plant which comprises a nucleic acid molecule according to claim 1 which is exogenous to said plant or plant cell.
42. A plant or cell of a plant which comprises a nucleic acid molecule according to claim 1, wherein said nucleic acid molecule is heterologous to said plant or said cell of a plant.
43. A plant or cell of a plant which has been transformed with a nucleic acid molecule according to claim 1.

- [illegible]

1054

SCHEMATIC 1

SCHEMATIC OF A GENE



ABSTRACT OF THE DISCLOSURE

The present invention provides DNA molecules that constitute fragments of the genome of a plant, and polypeptides encoded thereby. The DNA molecules are useful for specifying a gene product in cells, either as a promoter or as a protein coding sequence or as an UTR or as a 3' termination sequence, and are also useful in controlling the behavior of a gene in the chromosome, in controlling the expression of a gene or as tools for genetic mapping, recognizing or isolating identical or related DNA fragments, or identification of a particular individual organism, or for clustering of a group of organisms with a common trait.

5

Maximum Length Sequence corresponding to clone ID 101665

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 1
- Ceres seq_id 1481332

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 2
- Ceres seq_id 1481333
- Location of start within SEQ ID NO 1: at 203 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 107900

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 3
- Ceres seq_id 1481342

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 4
- Ceres seq_id 1481343
- Location of start within SEQ ID NO 3: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 5
- Ceres seq_id 1481344
- Location of start within SEQ ID NO 3: at 50 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 6
- Ceres seq_id 1481345
- Location of start within SEQ ID NO 3: at 518 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 108514

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 7
- Ceres seq_id 1481346

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 8
- Ceres seq_id 1481347
- Location of start within SEQ ID NO 7: at 629 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 9

- Ceres seq_id 1481348
- Location of start within SEQ ID NO 7: at 779 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 10
- Ceres seq_id 1481349
- Location of start within SEQ ID NO 7: at 828 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 109446

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 11
- Ceres seq_id 1481357

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 12
- Ceres seq_id 1481358
- Location of start within SEQ ID NO 11: at 342 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 13
- Ceres seq_id 1481359
- Location of start within SEQ ID NO 11: at 387 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 14
- Ceres seq_id 1481360
- Location of start within SEQ ID NO 11: at 396 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 113536

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 15
- Ceres seq_id 1481372

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 16
- Ceres seq_id 1481373
- Location of start within SEQ ID NO 15: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 17
- Ceres seq_id 1481374
- Location of start within SEQ ID NO 15: at 44 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 18
- Ceres seq_id 1481375
- Location of start within SEQ ID NO 15: at 348 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 115279

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 19
- Ceres seq_id 1481388

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 20
- Ceres seq_id 1481389
- Location of start within SEQ ID NO 19: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 21
- Ceres seq_id 1481390
- Location of start within SEQ ID NO 19: at 9 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 22
- Ceres seq_id 1481391
- Location of start within SEQ ID NO 19: at 63 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 118207

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 23
- Ceres seq_id 1481423

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 24
- Ceres seq_id 1481424
- Location of start within SEQ ID NO 23: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 25
- Ceres seq_id 1481425
- Location of start within SEQ ID NO 23: at 75 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 125028

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 26
- Ceres seq_id 1481471

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 27
- Ceres seq_id 1481472
- Location of start within SEQ ID NO 26: at 106 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 28
- Ceres seq_id 1481473
- Location of start within SEQ ID NO 26: at 169 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 29
- Ceres seq_id 1481474
- Location of start within SEQ ID NO 26: at 190 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 126108

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 30
- Ceres seq_id 1481479

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 31
- Ceres seq_id 1481480
- Location of start within SEQ ID NO 30: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 32

- Ceres seq_id 1481481
- Location of start within SEQ ID NO 30: at 114 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 33
- Ceres seq_id 1481482
- Location of start within SEQ ID NO 30: at 297 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 12613

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 34
- Ceres seq_id 1481483

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 35
- Ceres seq_id 1481484
- Location of start within SEQ ID NO 34: at 184 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 36
- Ceres seq_id 1481485
- Location of start within SEQ ID NO 34: at 268 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 37
- Ceres seq_id 1481486
- Location of start within SEQ ID NO 34: at 283 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 13607

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 38
- Ceres seq_id 1481487

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 39
- Ceres seq_id 1481488
- Location of start within SEQ ID NO 38: at 124 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 40
- Ceres seq_id 1481489
- Location of start within SEQ ID NO 38: at 133 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 41
- Ceres seq_id 1481490
- Location of start within SEQ ID NO 38: at 145 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 1367

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 42
- Ceres seq_id 1481491

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 43
- Ceres seq_id 1481492
- Location of start within SEQ ID NO 42: at 49 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 14568

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 44
- Ceres seq_id 1481504

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 45
- Ceres seq_id 1481505
- Location of start within SEQ ID NO 44: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 46
- Ceres seq_id 1481506
- Location of start within SEQ ID NO 44: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 47
- Ceres seq_id 1481507
- Location of start within SEQ ID NO 44: at 41 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 147980

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 48
- Ceres seq_id 1481516

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 49
- Ceres seq_id 1481517
- Location of start within SEQ ID NO 48: at 90 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 50
- Ceres seq_id 1481518
- Location of start within SEQ ID NO 48: at 186 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 51
- Ceres seq_id 1481519
- Location of start within SEQ ID NO 48: at 348 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 147983

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 52
- Ceres seq_id 1481520

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 53
- Ceres seq_id 1481521
- Location of start within SEQ ID NO 52: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 54
- Ceres seq_id 1481522
- Location of start within SEQ ID NO 52: at 68 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 55

- Ceres seq_id 1481523

- Location of start within SEQ ID NO 52: at 170 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 148070

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 56

- Ceres seq_id 1481524

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 57

- Ceres seq_id 1481525

- Location of start within SEQ ID NO 56: at 448 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 58

- Ceres seq_id 1481526

- Location of start within SEQ ID NO 56: at 1241 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 59

- Ceres seq_id 1481527

- Location of start within SEQ ID NO 56: at 1403 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 148232

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 60

- Ceres seq_id 1481532

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 61

- Ceres seq_id 1481533

- Location of start within SEQ ID NO 60: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 62

- Ceres seq_id 1481534

- Location of start within SEQ ID NO 60: at 108 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 63
- Ceres seq_id 1481535
- Location of start within SEQ ID NO 60: at 153 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 148887

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 64
- Ceres seq_id 1481540

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 65
- Ceres seq_id 1481541
- Location of start within SEQ ID NO 64: at 163 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 66
- Ceres seq_id 1481542
- Location of start within SEQ ID NO 64: at 220 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 67
- Ceres seq_id 1481543
- Location of start within SEQ ID NO 64: at 238 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 149204

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 68
- Ceres seq_id 1481544

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 69
- Ceres seq_id 1481545
- Location of start within SEQ ID NO 68: at 124 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 70
- Ceres seq_id 1481546
- Location of start within SEQ ID NO 68: at 178 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)
(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 71
 - Ceres seq_id 1481547
 - Location of start within SEQ ID NO 68: at 280 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)
(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 150293

- (A) Polynucleotide Sequence
- Pat. Appln. SEQ ID NO 72
 - Ceres seq_id 1481564
- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 73
 - Ceres seq_id 1481565
 - Location of start within SEQ ID NO 72: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)
(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 74
 - Ceres seq_id 1481566
 - Location of start within SEQ ID NO 72: at 60 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)
(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 75
 - Ceres seq_id 1481567
 - Location of start within SEQ ID NO 72: at 69 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)
(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 150540

- (A) Polynucleotide Sequence
- Pat. Appln. SEQ ID NO 76
 - Ceres seq_id 1481580
- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 77
 - Ceres seq_id 1481581
 - Location of start within SEQ ID NO 76: at 594 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)
(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 78

- Ceres seq_id 1481582

- Location of start within SEQ ID NO 76: at 630 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 79

- Ceres seq_id 1481583

- Location of start within SEQ ID NO 76: at 768 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 151413

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 80

- Ceres seq_id 1481596

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 81

- Ceres seq_id 1481597

- Location of start within SEQ ID NO 80: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 82

- Ceres seq_id 1481598

- Location of start within SEQ ID NO 80: at 87 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 83

- Ceres seq_id 1481599

- Location of start within SEQ ID NO 80: at 114 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 152305

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 84

- Ceres seq_id 1481613

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 85

- Ceres seq_id 1481614

- Location of start within SEQ ID NO 84: at 403 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 86
- Ceres seq_id 1481615
- Location of start within SEQ ID NO 84: at 562 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 87
- Ceres seq_id 1481616
- Location of start within SEQ ID NO 84: at 616 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 153154

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 88
- Ceres seq_id 1481621

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 89
- Ceres seq_id 1481622
- Location of start within SEQ ID NO 88: at 180 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 90
- Ceres seq_id 1481623
- Location of start within SEQ ID NO 88: at 291 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 91
- Ceres seq_id 1481624
- Location of start within SEQ ID NO 88: at 345 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 153808

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 92
- Ceres seq_id 1481625

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 93
- Ceres seq_id 1481626
- Location of start within SEQ ID NO 92: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 94
- Ceres seq_id 1481627
- Location of start within SEQ ID NO 92: at 88 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 95
- Ceres seq_id 1481628
- Location of start within SEQ ID NO 92: at 499 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 155661

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 96
- Ceres seq_id 1481632

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 97
- Ceres seq_id 1481633
- Location of start within SEQ ID NO 96: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 98
- Ceres seq_id 1481634
- Location of start within SEQ ID NO 96: at 9 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 155696

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 99
- Ceres seq_id 1481635

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 100
- Ceres seq_id 1481636
- Location of start within SEQ ID NO 99: at 152 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 101

- Ceres seq_id 1481637
- Location of start within SEQ ID NO 99: at 409 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 102
- Ceres seq_id 1481638
- Location of start within SEQ ID NO 99: at 457 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 155707

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 103
- Ceres seq_id 1481639

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 104
- Ceres seq_id 1481640
- Location of start within SEQ ID NO 103: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 105
- Ceres seq_id 1481641
- Location of start within SEQ ID NO 103: at 142 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 106
- Ceres seq_id 1481642
- Location of start within SEQ ID NO 103: at 712 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 156573

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 107
- Ceres seq_id 1481647

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 108
- Ceres seq_id 1481648
- Location of start within SEQ ID NO 107: at 156 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 109
- Ceres seq_id 1481649
- Location of start within SEQ ID NO 107: at 243 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 110
- Ceres seq_id 1481650
- Location of start within SEQ ID NO 107: at 429 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 1939

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 111
- Ceres seq_id 1481668

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 112
- Ceres seq_id 1481669
- Location of start within SEQ ID NO 111: at 201 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 113
- Ceres seq_id 1481670
- Location of start within SEQ ID NO 111: at 405 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 20783

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 114
- Ceres seq_id 1481681

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 115
- Ceres seq_id 1481682
- Location of start within SEQ ID NO 114: at 239 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 116
- Ceres seq_id 1481683
- Location of start within SEQ ID NO 114: at 398 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 218721

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 117
- Ceres seq_id 1481700

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 118
- Ceres seq_id 1481701
- Location of start within SEQ ID NO 117: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 119
- Ceres seq_id 1481702
- Location of start within SEQ ID NO 117: at 268 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 120
- Ceres seq_id 1481703
- Location of start within SEQ ID NO 117: at 292 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 218758

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 121
- Ceres seq_id 1481704

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 122
- Ceres seq_id 1481705
- Location of start within SEQ ID NO 121: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 123
- Ceres seq_id 1481706
- Location of start within SEQ ID NO 121: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 124

- Ceres seq_id 1481707
- Location of start within SEQ ID NO 121: at 60 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 220633

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 125
- Ceres seq_id 1481716

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 126
- Ceres seq_id 1481717
- Location of start within SEQ ID NO 125: at 55 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 127
- Ceres seq_id 1481718
- Location of start within SEQ ID NO 125: at 320 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 128
- Ceres seq_id 1481719
- Location of start within SEQ ID NO 125: at 395 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 220825

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 129
- Ceres seq_id 1481728

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 130
- Ceres seq_id 1481729
- Location of start within SEQ ID NO 129: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 131
- Ceres seq_id 1481730
- Location of start within SEQ ID NO 129: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 132
 - Ceres seq_id 1481731
 - Location of start within SEQ ID NO 129: at 214 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 220829

- (A) Polynucleotide Sequence
 - Pat. Appln. SEQ ID NO 133
 - Ceres seq_id 1481732
- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 134
 - Ceres seq_id 1481733
 - Location of start within SEQ ID NO 133: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 135
 - Ceres seq_id 1481734
 - Location of start within SEQ ID NO 133: at 169 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 136
 - Ceres seq_id 1481735
 - Location of start within SEQ ID NO 133: at 405 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 220846

- (A) Polynucleotide Sequence
 - Pat. Appln. SEQ ID NO 137
 - Ceres seq_id 1481740
- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 138
 - Ceres seq_id 1481741
 - Location of start within SEQ ID NO 137: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 139
 - Ceres seq_id 1481742
 - Location of start within SEQ ID NO 137: at 10 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 140
- Ceres seq_id 1481743
- Location of start within SEQ ID NO 137: at 22 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 220852

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 141
- Ceres seq_id 1481744

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 142
- Ceres seq_id 1481745
- Location of start within SEQ ID NO 141: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 143
- Ceres seq_id 1481746
- Location of start within SEQ ID NO 141: at 29 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 220854

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 144
- Ceres seq_id 1481747

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 145
- Ceres seq_id 1481748
- Location of start within SEQ ID NO 144: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 146
- Ceres seq_id 1481749
- Location of start within SEQ ID NO 144: at 74 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 147

- Ceres seq_id 1481750
- Location of start within SEQ ID NO 144: at 95 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 220915

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 148
- Ceres seq_id 1481755

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 149
- Ceres seq_id 1481756
- Location of start within SEQ ID NO 148: at 178 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 220934

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 150
- Ceres seq_id 1481764

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 151
- Ceres seq_id 1481765
- Location of start within SEQ ID NO 150: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 152
- Ceres seq_id 1481766
- Location of start within SEQ ID NO 150: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 220944

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 153
- Ceres seq_id 1481770

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 154
- Ceres seq_id 1481771
- Location of start within SEQ ID NO 153: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 155
- Ceres seq_id 1481772
- Location of start within SEQ ID NO 153: at 96 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 226475

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 156
- Ceres seq_id 1481775

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 157
- Ceres seq_id 1481776
- Location of start within SEQ ID NO 156: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 158
- Ceres seq_id 1481777
- Location of start within SEQ ID NO 156: at 68 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 159
- Ceres seq_id 1481778
- Location of start within SEQ ID NO 156: at 255 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 226483

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 160
- Ceres seq_id 1481779

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 161
- Ceres seq_id 1481780
- Location of start within SEQ ID NO 160: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 226501

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 162
- Ceres seq_id 1481789

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 163
- Ceres seq_id 1481790
- Location of start within SEQ ID NO 162: at 109 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 164
- Ceres seq_id 1481791
- Location of start within SEQ ID NO 162: at 229 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 226516

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 165
- Ceres seq_id 1481792

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 166
- Ceres seq_id 1481793
- Location of start within SEQ ID NO 165: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 167
- Ceres seq_id 1481794
- Location of start within SEQ ID NO 165: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 168
- Ceres seq_id 1481795
- Location of start within SEQ ID NO 165: at 67 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 227154

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 169
- Ceres seq_id 1481796

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 170
- Ceres seq_id 1481797
- Location of start within SEQ ID NO 169: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 171

- Ceres seq_id 1481798
- Location of start within SEQ ID NO 169: at 118 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 227202

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 172
- Ceres seq_id 1481799

Maximum Length Sequence corresponding to clone ID 227468

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 173
- Ceres seq_id 1481800

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 174
- Ceres seq_id 1481801
- Location of start within SEQ ID NO 173: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 175
- Ceres seq_id 1481802
- Location of start within SEQ ID NO 173: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 176
- Ceres seq_id 1481803
- Location of start within SEQ ID NO 173: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 227480

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 177
- Ceres seq_id 1481808

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 178
- Ceres seq_id 1481809
- Location of start within SEQ ID NO 177: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 179
- Ceres seq_id 1481810
- Location of start within SEQ ID NO 177: at 15 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 180
- Ceres seq_id 1481811
- Location of start within SEQ ID NO 177: at 45 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 227719

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 181
- Ceres seq_id 1481815

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 182
- Ceres seq_id 1481816
- Location of start within SEQ ID NO 181: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 183
- Ceres seq_id 1481817
- Location of start within SEQ ID NO 181: at 40 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 184
- Ceres seq_id 1481818
- Location of start within SEQ ID NO 181: at 106 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 227812

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 185
- Ceres seq_id 1481819

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 186
- Ceres seq_id 1481820
- Location of start within SEQ ID NO 185: at 51 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 187
- Ceres seq_id 1481821
- Location of start within SEQ ID NO 185: at 57 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 188
- Ceres seq_id 1481822
- Location of start within SEQ ID NO 185: at 66 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 227814

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 189
- Ceres seq_id 1481823

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 190
- Ceres seq_id 1481824
- Location of start within SEQ ID NO 189: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 191
- Ceres seq_id 1481825
- Location of start within SEQ ID NO 189: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 192
- Ceres seq_id 1481826
- Location of start within SEQ ID NO 189: at 70 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 227825

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 193
- Ceres seq_id 1481827

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 194
- Ceres seq_id 1481828
- Location of start within SEQ ID NO 193: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 195
 - Ceres seq_id 1481829
 - Location of start within SEQ ID NO 193: at 167 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 196
 - Ceres seq_id 1481830
 - Location of start within SEQ ID NO 193: at 215 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 229883

- (A) Polynucleotide Sequence
- Pat. Appln. SEQ ID NO 197
 - Ceres seq_id 1481831
- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 198
 - Ceres seq_id 1481832
 - Location of start within SEQ ID NO 197: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 199
 - Ceres seq_id 1481833
 - Location of start within SEQ ID NO 197: at 69 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 200
 - Ceres seq_id 1481834
 - Location of start within SEQ ID NO 197: at 96 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 231825

- (A) Polynucleotide Sequence
- Pat. Appln. SEQ ID NO 201
 - Ceres seq_id 1481839
- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 202
 - Ceres seq_id 1481840
 - Location of start within SEQ ID NO 201: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 203
- Ceres seq_id 1481841
- Location of start within SEQ ID NO 201: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 204
- Ceres seq_id 1481842
- Location of start within SEQ ID NO 201: at 272 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 232410

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 205
- Ceres seq_id 1481847

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 206
- Ceres seq_id 1481848
- Location of start within SEQ ID NO 205: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 207
- Ceres seq_id 1481849
- Location of start within SEQ ID NO 205: at 44 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 208
- Ceres seq_id 1481850
- Location of start within SEQ ID NO 205: at 128 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 232492

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 209
- Ceres seq_id 1481851

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 210
- Ceres seq_id 1481852
- Location of start within SEQ ID NO 209: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 211
- Ceres seq_id 1481853
- Location of start within SEQ ID NO 209: at 62 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 212
- Ceres seq_id 1481854
- Location of start within SEQ ID NO 209: at 122 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 237301

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 213
- Ceres seq_id 1481859

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 214
- Ceres seq_id 1481860
- Location of start within SEQ ID NO 213: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 215
- Ceres seq_id 1481861
- Location of start within SEQ ID NO 213: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 216
- Ceres seq_id 1481862
- Location of start within SEQ ID NO 213: at 5 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 237328

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 217
- Ceres seq_id 1481863
- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 218
 - Ceres seq_id 1481864
 - Location of start within SEQ ID NO 217: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 219
 - Ceres seq_id 1481865
 - Location of start within SEQ ID NO 217: at 71 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 246496

- (A) Polynucleotide Sequence
 - Pat. Appln. SEQ ID NO 220
 - Ceres seq_id 1481873
- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 221
 - Ceres seq_id 1481874
 - Location of start within SEQ ID NO 220: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 222
 - Ceres seq_id 1481875
 - Location of start within SEQ ID NO 220: at 379 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 246936

- (A) Polynucleotide Sequence
 - Pat. Appln. SEQ ID NO 223
 - Ceres seq_id 1481885
- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 224
 - Ceres seq_id 1481886
 - Location of start within SEQ ID NO 223: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 225
 - Ceres seq_id 1481887
 - Location of start within SEQ ID NO 223: at 48 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 226
- Ceres seq_id 1481888
- Location of start within SEQ ID NO 223: at 109 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 247196

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 227
- Ceres seq_id 1481893

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 228
- Ceres seq_id 1481894
- Location of start within SEQ ID NO 227: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 229
- Ceres seq_id 1481895
- Location of start within SEQ ID NO 227: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 230
- Ceres seq_id 1481896
- Location of start within SEQ ID NO 227: at 271 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 247299

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 231
- Ceres seq_id 1481897

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 232
- Ceres seq_id 1481898
- Location of start within SEQ ID NO 231: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 233
- Ceres seq_id 1481899
- Location of start within SEQ ID NO 231: at 37 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 234
- Ceres seq_id 1481900
- Location of start within SEQ ID NO 231: at 70 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 250561

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 235
- Ceres seq_id 1481901

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 236
- Ceres seq_id 1481902
- Location of start within SEQ ID NO 235: at 150 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 250647

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 237
- Ceres seq_id 1481903

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 238
- Ceres seq_id 1481904
- Location of start within SEQ ID NO 237: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 239
- Ceres seq_id 1481905
- Location of start within SEQ ID NO 237: at 68 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 240
- Ceres seq_id 1481906
- Location of start within SEQ ID NO 237: at 116 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

Maximum Length Sequence corresponding to clone ID 250663

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 241
- Ceres seq_id 1481907

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 242
- Ceres seq_id 1481908
- Location of start within SEQ ID NO 241: at 165 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 250775

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 243
- Ceres seq_id 1481913

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 244
- Ceres seq_id 1481914
- Location of start within SEQ ID NO 243: at 126 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 245
- Ceres seq_id 1481915
- Location of start within SEQ ID NO 243: at 291 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 251921

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 246
- Ceres seq_id 1481916

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 247
- Ceres seq_id 1481917
- Location of start within SEQ ID NO 246: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 248
- Ceres seq_id 1481918
- Location of start within SEQ ID NO 246: at 231 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 252000

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 249
- Ceres seq_id 1481919

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 250
- Ceres seq_id 1481920
- Location of start within SEQ ID NO 249: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 251
- Ceres seq_id 1481921
- Location of start within SEQ ID NO 249: at 48 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 252
- Ceres seq_id 1481922
- Location of start within SEQ ID NO 249: at 198 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 252002

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 253
- Ceres seq_id 1481923

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 254
- Ceres seq_id 1481924
- Location of start within SEQ ID NO 253: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 255
- Ceres seq_id 1481925
- Location of start within SEQ ID NO 253: at 424 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 257043

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 256
- Ceres seq_id 1481941

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 257

- Ceres seq_id 1481942
- Location of start within SEQ ID NO 256: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 258
- Ceres seq_id 1481943
- Location of start within SEQ ID NO 256: at 71 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 259
- Ceres seq_id 1481944
- Location of start within SEQ ID NO 256: at 74 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 257207

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 260
- Ceres seq_id 1481949

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 261
- Ceres seq_id 1481950
- Location of start within SEQ ID NO 260: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 262
- Ceres seq_id 1481951
- Location of start within SEQ ID NO 260: at 276 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 263
- Ceres seq_id 1481952
- Location of start within SEQ ID NO 260: at 454 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 265955

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 264

- Ceres seq_id 1481965

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 265
- Ceres seq_id 1481966
- Location of start within SEQ ID NO 264: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 266
- Ceres seq_id 1481967
- Location of start within SEQ ID NO 264: at 103 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 267
- Ceres seq_id 1481968
- Location of start within SEQ ID NO 264: at 327 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 266374

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 268
- Ceres seq_id 1481973

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 269
- Ceres seq_id 1481974
- Location of start within SEQ ID NO 268: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 270
- Ceres seq_id 1481975
- Location of start within SEQ ID NO 268: at 113 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 266934

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 271
- Ceres seq_id 1481976

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 272
- Ceres seq_id 1481977
- Location of start within SEQ ID NO 271: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 273
- Ceres seq_id 1481978
- Location of start within SEQ ID NO 271: at 5 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 274
- Ceres seq_id 1481979
- Location of start within SEQ ID NO 271: at 65 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 266951

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 275
- Ceres seq_id 1481980

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 276
- Ceres seq_id 1481981
- Location of start within SEQ ID NO 275: at 54 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 277
- Ceres seq_id 1481982
- Location of start within SEQ ID NO 275: at 307 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 267031

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 278
- Ceres seq_id 1481983

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 279
- Ceres seq_id 1481984
- Location of start within SEQ ID NO 278: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 280

- Ceres seq_id 1481985

- Location of start within SEQ ID NO 278: at 31 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 267032

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 281

- Ceres seq_id 1481986

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 282

- Ceres seq_id 1481987

- Location of start within SEQ ID NO 281: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 283

- Ceres seq_id 1481988

- Location of start within SEQ ID NO 281: at 38 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 284

- Ceres seq_id 1481989

- Location of start within SEQ ID NO 281: at 131 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 267296

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 285

- Ceres seq_id 1481990

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 286

- Ceres seq_id 1481991

- Location of start within SEQ ID NO 285: at 157 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 287

- Ceres seq_id 1481992

- Location of start within SEQ ID NO 285: at 163 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 288
 - Ceres seq_id 1481993
 - Location of start within SEQ ID NO 285: at 412 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 267626

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 289
- Ceres seq_id 1481994

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 290
- Ceres seq_id 1481995
- Location of start within SEQ ID NO 289: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 291
- Ceres seq_id 1481996
- Location of start within SEQ ID NO 289: at 157 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 292
- Ceres seq_id 1481997
- Location of start within SEQ ID NO 289: at 175 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 268353

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 293
- Ceres seq_id 1482009

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 294
- Ceres seq_id 1482010
- Location of start within SEQ ID NO 293: at 72 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 295
- Ceres seq_id 1482011
- Location of start within SEQ ID NO 293: at 144 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 296
- Ceres seq_id 1482012
- Location of start within SEQ ID NO 293: at 321 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 268652

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 297
- Ceres seq_id 1482013

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 298
- Ceres seq_id 1482014
- Location of start within SEQ ID NO 297: at 33 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 299
- Ceres seq_id 1482015
- Location of start within SEQ ID NO 297: at 156 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 268680

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 300
- Ceres seq_id 1482016

Maximum Length Sequence corresponding to clone ID 269248

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 301
- Ceres seq_id 1482021

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 302
- Ceres seq_id 1482022
- Location of start within SEQ ID NO 301: at 175 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 303
- Ceres seq_id 1482023
- Location of start within SEQ ID NO 301: at 190 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 304
 - Ceres seq_id 1482024
 - Location of start within SEQ ID NO 301: at 262 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 270513

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 305
- Ceres seq_id 1482029

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 306
- Ceres seq_id 1482030
- Location of start within SEQ ID NO 305: at 86 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 307
- Ceres seq_id 1482031
- Location of start within SEQ ID NO 305: at 194 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 308
- Ceres seq_id 1482032
- Location of start within SEQ ID NO 305: at 203 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 270518

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 309
- Ceres seq_id 1482033

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 310
- Ceres seq_id 1482034
- Location of start within SEQ ID NO 309: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 311
- Ceres seq_id 1482035
- Location of start within SEQ ID NO 309: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 312
- Ceres seq_id 1482036
- Location of start within SEQ ID NO 309: at 119 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 271717

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 313
- Ceres seq_id 1482041

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 314
- Ceres seq_id 1482042
- Location of start within SEQ ID NO 313: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 315
- Ceres seq_id 1482043
- Location of start within SEQ ID NO 313: at 491 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 316
- Ceres seq_id 1482044
- Location of start within SEQ ID NO 313: at 518 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 271756

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 317
- Ceres seq_id 1482045

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 318
- Ceres seq_id 1482046
- Location of start within SEQ ID NO 317: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 319
- Ceres seq_id 1482047
- Location of start within SEQ ID NO 317: at 71 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 320
- Ceres seq_id 1482048
- Location of start within SEQ ID NO 317: at 149 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 271765

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 321
- Ceres seq_id 1482049

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 322
- Ceres seq_id 1482050
- Location of start within SEQ ID NO 321: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 323
- Ceres seq_id 1482051
- Location of start within SEQ ID NO 321: at 27 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 324
- Ceres seq_id 1482052
- Location of start within SEQ ID NO 321: at 45 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 271936

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 325
- Ceres seq_id 1482053

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 326
- Ceres seq_id 1482054
- Location of start within SEQ ID NO 325: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

Maximum Length Sequence corresponding to clone ID 272121

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 327
- Ceres seq_id 1482066

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 328
- Ceres seq_id 1482067
- Location of start within SEQ ID NO 327: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 329
- Ceres seq_id 1482068
- Location of start within SEQ ID NO 327: at 113 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272124

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 330
- Ceres seq_id 1482069

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 331
- Ceres seq_id 1482070
- Location of start within SEQ ID NO 330: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 332
- Ceres seq_id 1482071
- Location of start within SEQ ID NO 330: at 162 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 333
- Ceres seq_id 1482072
- Location of start within SEQ ID NO 330: at 186 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272142

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 334
- Ceres seq_id 1482073

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 335
- Ceres seq_id 1482074
- Location of start within SEQ ID NO 334: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 336
- Ceres seq_id 1482075
- Location of start within SEQ ID NO 334: at 272 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 337
- Ceres seq_id 1482076
- Location of start within SEQ ID NO 334: at 344 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272155

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 338
- Ceres seq_id 1482081

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 339
- Ceres seq_id 1482082
- Location of start within SEQ ID NO 338: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 340
- Ceres seq_id 1482083
- Location of start within SEQ ID NO 338: at 8 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 341
- Ceres seq_id 1482084
- Location of start within SEQ ID NO 338: at 178 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272156

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 342
- Ceres seq_id 1482085

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 343
- Ceres seq_id 1482086
- Location of start within SEQ ID NO 342: at 302 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272162

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 344
- Ceres seq_id 1482091

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 345
- Ceres seq_id 1482092
- Location of start within SEQ ID NO 344: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 346
- Ceres seq_id 1482093
- Location of start within SEQ ID NO 344: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 347
- Ceres seq_id 1482094
- Location of start within SEQ ID NO 344: at 102 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272166

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 348
- Ceres seq_id 1482095

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 349
- Ceres seq_id 1482096
- Location of start within SEQ ID NO 348: at 229 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 350
- Ceres seq_id 1482097

- Location of start within SEQ ID NO 348: at 322 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272200

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 351

- Ceres seq_id 1482102

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 352

- Ceres seq_id 1482103

- Location of start within SEQ ID NO 351: at 282 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 353

- Ceres seq_id 1482104

- Location of start within SEQ ID NO 351: at 309 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 354

- Ceres seq_id 1482105

- Location of start within SEQ ID NO 351: at 366 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272214

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 355

- Ceres seq_id 1482106

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 356

- Ceres seq_id 1482107

- Location of start within SEQ ID NO 355: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 357

- Ceres seq_id 1482108

- Location of start within SEQ ID NO 355: at 240 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272239

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 358
- Ceres seq_id 1482113

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 359
- Ceres seq_id 1482114
- Location of start within SEQ ID NO 358: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 360
- Ceres seq_id 1482115
- Location of start within SEQ ID NO 358: at 97 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 361
- Ceres seq_id 1482116
- Location of start within SEQ ID NO 358: at 121 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272250

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 362
- Ceres seq_id 1482117

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 363
- Ceres seq_id 1482118
- Location of start within SEQ ID NO 362: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 364
- Ceres seq_id 1482119
- Location of start within SEQ ID NO 362: at 69 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 365
- Ceres seq_id 1482120
- Location of start within SEQ ID NO 362: at 264 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272258

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 366
- Ceres seq_id 1482121

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 367
- Ceres seq_id 1482122
- Location of start within SEQ ID NO 366: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272301

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 368
- Ceres seq_id 1482127

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 369
- Ceres seq_id 1482128
- Location of start within SEQ ID NO 368: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 370
- Ceres seq_id 1482129
- Location of start within SEQ ID NO 368: at 92 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 371
- Ceres seq_id 1482130
- Location of start within SEQ ID NO 368: at 529 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272312

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 372
- Ceres seq_id 1482131

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 373
- Ceres seq_id 1482132
- Location of start within SEQ ID NO 372: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 374
 - Ceres seq_id 1482133
 - Location of start within SEQ ID NO 372: at 42 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 375
 - Ceres seq_id 1482134
 - Location of start within SEQ ID NO 372: at 168 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272389

- (A) Polynucleotide Sequence
- Pat. Appln. SEQ ID NO 376
 - Ceres seq_id 1482135
- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 377
 - Ceres seq_id 1482136
 - Location of start within SEQ ID NO 376: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 378
 - Ceres seq_id 1482137
 - Location of start within SEQ ID NO 376: at 7 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272410

- (A) Polynucleotide Sequence
- Pat. Appln. SEQ ID NO 379
 - Ceres seq_id 1482142
- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 380
 - Ceres seq_id 1482143
 - Location of start within SEQ ID NO 379: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 381
 - Ceres seq_id 1482144
 - Location of start within SEQ ID NO 379: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 382
- Ceres seq_id 1482145
- Location of start within SEQ ID NO 379: at 92 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272459

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 383
- Ceres seq_id 1482153

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 384
- Ceres seq_id 1482154
- Location of start within SEQ ID NO 383: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 385
- Ceres seq_id 1482155
- Location of start within SEQ ID NO 383: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 386
- Ceres seq_id 1482156
- Location of start within SEQ ID NO 383: at 326 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272486

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 387
- Ceres seq_id 1482157

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 388
- Ceres seq_id 1482158
- Location of start within SEQ ID NO 387: at 139 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 389
- Ceres seq_id 1482159
- Location of start within SEQ ID NO 387: at 301 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 272506

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 390
- Ceres seq_id 1482164

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 391
- Ceres seq_id 1482165
- Location of start within SEQ ID NO 390: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 392
- Ceres seq_id 1482166
- Location of start within SEQ ID NO 390: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 275387

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 393
- Ceres seq_id 1482167

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 394
- Ceres seq_id 1482168
- Location of start within SEQ ID NO 393: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 275402

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 395
- Ceres seq_id 1482169

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 396
- Ceres seq_id 1482170
- Location of start within SEQ ID NO 395: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 397
- Ceres seq_id 1482171

- Location of start within SEQ ID NO 395: at 79 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 398

- Ceres seq_id 1482172

- Location of start within SEQ ID NO 395: at 360 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 275778

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 399

- Ceres seq_id 1482177

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 400

- Ceres seq_id 1482178

- Location of start within SEQ ID NO 399: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 401

- Ceres seq_id 1482179

- Location of start within SEQ ID NO 399: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 402

- Ceres seq_id 1482180

- Location of start within SEQ ID NO 399: at 277 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 275803

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 403

- Ceres seq_id 1482188

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 404

- Ceres seq_id 1482189

- Location of start within SEQ ID NO 403: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Attorney Docket No. 2750-1096P

Client Docket No. 80142.004

Maximum Length Sequence corresponding to clone ID 276193

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 405
- Ceres seq_id 1482193

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 406
- Ceres seq_id 1482194
- Location of start within SEQ ID NO 405: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 294676

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 407
- Ceres seq_id 1482205

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 408
- Ceres seq_id 1482206
- Location of start within SEQ ID NO 407: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 296069

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 409
- Ceres seq_id 1482207

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 410
- Ceres seq_id 1482208
- Location of start within SEQ ID NO 409: at 124 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 411
- Ceres seq_id 1482209
- Location of start within SEQ ID NO 409: at 226 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 412
- Ceres seq_id 1482210
- Location of start within SEQ ID NO 409: at 271 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 296091

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 413
- Ceres seq_id 1482217
- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 414
 - Ceres seq_id 1482218
 - Location of start within SEQ ID NO 413: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 296096

- (A) Polynucleotide Sequence
 - Pat. Appln. SEQ ID NO 415
 - Ceres seq_id 1482219
- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 416
 - Ceres seq_id 1482220
 - Location of start within SEQ ID NO 415: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 417
 - Ceres seq_id 1482221
 - Location of start within SEQ ID NO 415: at 72 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 296205

- (A) Polynucleotide Sequence
 - Pat. Appln. SEQ ID NO 418
 - Ceres seq_id 1482230
- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 419
 - Ceres seq_id 1482231
 - Location of start within SEQ ID NO 418: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 420
 - Ceres seq_id 1482232
 - Location of start within SEQ ID NO 418: at 125 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 421
 - Ceres seq_id 1482233
 - Location of start within SEQ ID NO 418: at 152 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 296209

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 422
- Ceres seq_id 1482234

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 423
- Ceres seq_id 1482235
- Location of start within SEQ ID NO 422: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 424
- Ceres seq_id 1482236
- Location of start within SEQ ID NO 422: at 221 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 425
- Ceres seq_id 1482237
- Location of start within SEQ ID NO 422: at 287 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 296211

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 426
- Ceres seq_id 1482238

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 427
- Ceres seq_id 1482239
- Location of start within SEQ ID NO 426: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 428
- Ceres seq_id 1482240
- Location of start within SEQ ID NO 426: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 296215

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 429
- Ceres seq_id 1482245

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 430
- Ceres seq_id 1482246
- Location of start within SEQ ID NO 429: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 431
- Ceres seq_id 1482247
- Location of start within SEQ ID NO 429: at 176 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 296228

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 432
- Ceres seq_id 1482248

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 433
- Ceres seq_id 1482249
- Location of start within SEQ ID NO 432: at 120 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 434
- Ceres seq_id 1482250
- Location of start within SEQ ID NO 432: at 249 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 435
- Ceres seq_id 1482251
- Location of start within SEQ ID NO 432: at 312 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 296237

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 436
- Ceres seq_id 1482254

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 437
- Ceres seq_id 1482255

- Location of start within SEQ ID NO 436: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 438
- Ceres seq_id 1482256
- Location of start within SEQ ID NO 436: at 58 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 296246

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 439
- Ceres seq_id 1482257

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 440
- Ceres seq_id 1482258
- Location of start within SEQ ID NO 439: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 441
- Ceres seq_id 1482259
- Location of start within SEQ ID NO 439: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 442
- Ceres seq_id 1482260
- Location of start within SEQ ID NO 439: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 296620

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 443
- Ceres seq_id 1482261

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 444
- Ceres seq_id 1482262
- Location of start within SEQ ID NO 443: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 445
- Ceres seq_id 1482263
- Location of start within SEQ ID NO 443: at 313 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 296648

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 446
- Ceres seq_id 1482264

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 447
- Ceres seq_id 1482265
- Location of start within SEQ ID NO 446: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 297691

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 448
- Ceres seq_id 1482270

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 449
- Ceres seq_id 1482271
- Location of start within SEQ ID NO 448: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 450
- Ceres seq_id 1482272
- Location of start within SEQ ID NO 448: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 451
- Ceres seq_id 1482273
- Location of start within SEQ ID NO 448: at 199 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 297711

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 452
- Ceres seq_id 1482274

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 453

- Ceres seq_id 1482275
- Location of start within SEQ ID NO 452: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 454
- Ceres seq_id 1482276
- Location of start within SEQ ID NO 452: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 455
- Ceres seq_id 1482277
- Location of start within SEQ ID NO 452: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 299123

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 456
- Ceres seq_id 1482282

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 457
- Ceres seq_id 1482283
- Location of start within SEQ ID NO 456: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 458
- Ceres seq_id 1482284
- Location of start within SEQ ID NO 456: at 223 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 459
- Ceres seq_id 1482285
- Location of start within SEQ ID NO 456: at 286 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 299990

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 460

- Ceres seq_id 1482289

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 461
- Ceres seq_id 1482290
- Location of start within SEQ ID NO 460: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 462
- Ceres seq_id 1482291
- Location of start within SEQ ID NO 460: at 21 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 463
- Ceres seq_id 1482292
- Location of start within SEQ ID NO 460: at 123 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 299991

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 464
- Ceres seq_id 1482293

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 465
- Ceres seq_id 1482294
- Location of start within SEQ ID NO 464: at 184 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 466
- Ceres seq_id 1482295
- Location of start within SEQ ID NO 464: at 226 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 467
- Ceres seq_id 1482296
- Location of start within SEQ ID NO 464: at 349 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Attorney Docket No. 2750-1096P

Client Docket No. 80142.004

Table 1

Page 61

Maximum Length Sequence corresponding to clone ID 300985

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 468

- Ceres seq_id 1482297

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 469

- Ceres seq_id 1482298

- Location of start within SEQ ID NO 468: at 72 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 300986

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 470

- Ceres seq_id 1482299

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 471

- Ceres seq_id 1482300

- Location of start within SEQ ID NO 470: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 472

- Ceres seq_id 1482301

- Location of start within SEQ ID NO 470: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 300987

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 473

- Ceres seq_id 1482302

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 474

- Ceres seq_id 1482303

- Location of start within SEQ ID NO 473: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301009

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 475

- Ceres seq_id 1482307

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 476

- Ceres seq_id 1482308

- Location of start within SEQ ID NO 475: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 477
 - Ceres seq_id 1482309
 - Location of start within SEQ ID NO 475: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301084

- (A) Polynucleotide Sequence
- Pat. Appln. SEQ ID NO 478
 - Ceres seq_id 1482322
- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 479
 - Ceres seq_id 1482323
 - Location of start within SEQ ID NO 478: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 480
 - Ceres seq_id 1482324
 - Location of start within SEQ ID NO 478: at 86 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 481
 - Ceres seq_id 1482325
 - Location of start within SEQ ID NO 478: at 319 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301128

- (A) Polynucleotide Sequence
- Pat. Appln. SEQ ID NO 482
 - Ceres seq_id 1482334
- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 483
 - Ceres seq_id 1482335
 - Location of start within SEQ ID NO 482: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301143

- (A) Polynucleotide Sequence
- Pat. Appln. SEQ ID NO 484
 - Ceres seq_id 1482336

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 485
- Ceres seq_id 1482337
- Location of start within SEQ ID NO 484: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 486
- Ceres seq_id 1482338
- Location of start within SEQ ID NO 484: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301452

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 487
- Ceres seq_id 1482339

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 488
- Ceres seq_id 1482340
- Location of start within SEQ ID NO 487: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 489
- Ceres seq_id 1482341
- Location of start within SEQ ID NO 487: at 96 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 490
- Ceres seq_id 1482342
- Location of start within SEQ ID NO 487: at 138 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301456

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 491
- Ceres seq_id 1482346

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 492
- Ceres seq_id 1482347
- Location of start within SEQ ID NO 491: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 493
- Ceres seq_id 1482348
- Location of start within SEQ ID NO 491: at 120 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301464

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 494
- Ceres seq_id 1482349

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 495
- Ceres seq_id 1482350
- Location of start within SEQ ID NO 494: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 496
- Ceres seq_id 1482351
- Location of start within SEQ ID NO 494: at 135 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 497
- Ceres seq_id 1482352
- Location of start within SEQ ID NO 494: at 195 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301481

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 498
- Ceres seq_id 1482353

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 499
- Ceres seq_id 1482354
- Location of start within SEQ ID NO 498: at 98 nt.

(C) Nomination and Annotation of Domains within Predicted

Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 500

- Ceres seq_id 1482355
- Location of start within SEQ ID NO 498: at 242 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301483

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 501
- Ceres seq_id 1482356

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 502
- Ceres seq_id 1482357
- Location of start within SEQ ID NO 501: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 503
- Ceres seq_id 1482358
- Location of start within SEQ ID NO 501: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301504

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 504
- Ceres seq_id 1482359

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 505
- Ceres seq_id 1482360
- Location of start within SEQ ID NO 504: at 14 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 506
- Ceres seq_id 1482361
- Location of start within SEQ ID NO 504: at 294 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 507
- Ceres seq_id 1482362
- Location of start within SEQ ID NO 504: at 297 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301535

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 508
- Ceres seq_id 1482363

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 509
- Ceres seq_id 1482364
- Location of start within SEQ ID NO 508: at 51 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 510
- Ceres seq_id 1482365
- Location of start within SEQ ID NO 508: at 86 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 511
- Ceres seq_id 1482366
- Location of start within SEQ ID NO 508: at 205 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301541

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 512
- Ceres seq_id 1482371

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 513
- Ceres seq_id 1482372
- Location of start within SEQ ID NO 512: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 514
- Ceres seq_id 1482373
- Location of start within SEQ ID NO 512: at 206 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 515
- Ceres seq_id 1482374
- Location of start within SEQ ID NO 512: at 224 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301552

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 516
- Ceres seq_id 1482375

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 517
- Ceres seq_id 1482376
- Location of start within SEQ ID NO 516: at 151 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 518
- Ceres seq_id 1482377
- Location of start within SEQ ID NO 516: at 166 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301559

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 519
- Ceres seq_id 1482378

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 520
- Ceres seq_id 1482379
- Location of start within SEQ ID NO 519: at 83 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 521
- Ceres seq_id 1482380
- Location of start within SEQ ID NO 519: at 113 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 522
- Ceres seq_id 1482381
- Location of start within SEQ ID NO 519: at 143 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301584

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 523
- Ceres seq_id 1482382
- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 524
 - Ceres seq_id 1482383
 - Location of start within SEQ ID NO 523: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 525
 - Ceres seq_id 1482384
 - Location of start within SEQ ID NO 523: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 526
 - Ceres seq_id 1482385
 - Location of start within SEQ ID NO 523: at 191 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301586

- (A) Polynucleotide Sequence
 - Pat. Appln. SEQ ID NO 527
 - Ceres seq_id 1482386
- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 528
 - Ceres seq_id 1482387
 - Location of start within SEQ ID NO 527: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 529
 - Ceres seq_id 1482388
 - Location of start within SEQ ID NO 527: at 34 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
 - Pat. Appln. SEQ ID NO 530
 - Ceres seq_id 1482389
 - Location of start within SEQ ID NO 527: at 437 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301930

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 531
- Ceres seq_id 1482398

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 532
- Ceres seq_id 1482399
- Location of start within SEQ ID NO 531: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 533
- Ceres seq_id 1482400
- Location of start within SEQ ID NO 531: at 98 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 534
- Ceres seq_id 1482401
- Location of start within SEQ ID NO 531: at 425 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301956

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 535
- Ceres seq_id 1482402

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 536
- Ceres seq_id 1482403
- Location of start within SEQ ID NO 535: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301961

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 537
- Ceres seq_id 1482404

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 538
- Ceres seq_id 1482405
- Location of start within SEQ ID NO 537: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 539
- Ceres seq_id 1482406
- Location of start within SEQ ID NO 537: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 540
- Ceres seq_id 1482407
- Location of start within SEQ ID NO 537: at 182 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301981

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 541
- Ceres seq_id 1482408

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 542
- Ceres seq_id 1482409
- Location of start within SEQ ID NO 541: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 543
- Ceres seq_id 1482410
- Location of start within SEQ ID NO 541: at 442 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 301994

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 544
- Ceres seq_id 1482411

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 545
- Ceres seq_id 1482412
- Location of start within SEQ ID NO 544: at 94 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 546
- Ceres seq_id 1482413
- Location of start within SEQ ID NO 544: at 155 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 547
 - Ceres seq_id 1482414
 - Location of start within SEQ ID NO 544: at 620 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 302016

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 548
- Ceres seq_id 1482415

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 549
- Ceres seq_id 1482416
- Location of start within SEQ ID NO 548: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 550
- Ceres seq_id 1482417
- Location of start within SEQ ID NO 548: at 107 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 302030

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 551
- Ceres seq_id 1482418

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 552
- Ceres seq_id 1482419
- Location of start within SEQ ID NO 551: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 553
- Ceres seq_id 1482420
- Location of start within SEQ ID NO 551: at 105 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 554
- Ceres seq_id 1482421
- Location of start within SEQ ID NO 551: at 135 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 302415

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 555
- Ceres seq_id 1482422

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 556
- Ceres seq_id 1482423
- Location of start within SEQ ID NO 555: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 304700

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 557
- Ceres seq_id 1482424

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 558
- Ceres seq_id 1482425
- Location of start within SEQ ID NO 557: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 559
- Ceres seq_id 1482426
- Location of start within SEQ ID NO 557: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 560
- Ceres seq_id 1482427
- Location of start within SEQ ID NO 557: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 304743

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 561
- Ceres seq_id 1482428

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 562
- Ceres seq_id 1482429
- Location of start within SEQ ID NO 561: at 206 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)
(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 304764

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 563
- Ceres seq_id 1482430

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 564
- Ceres seq_id 1482431
- Location of start within SEQ ID NO 563: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 565
- Ceres seq_id 1482432
- Location of start within SEQ ID NO 563: at 11 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 566
- Ceres seq_id 1482433
- Location of start within SEQ ID NO 563: at 86 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 304769

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 567
- Ceres seq_id 1482434

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 568
- Ceres seq_id 1482435
- Location of start within SEQ ID NO 567: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 569
- Ceres seq_id 1482436
- Location of start within SEQ ID NO 567: at 9 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 570

- Ceres seq_id 1482437
- Location of start within SEQ ID NO 567: at 18 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 305124

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 571
- Ceres seq_id 1482438

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 572
- Ceres seq_id 1482439
- Location of start within SEQ ID NO 571: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 30994

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 573
- Ceres seq_id 1482444

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 574
- Ceres seq_id 1482445
- Location of start within SEQ ID NO 573: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 575
- Ceres seq_id 1482446
- Location of start within SEQ ID NO 573: at 134 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 576
- Ceres seq_id 1482447
- Location of start within SEQ ID NO 573: at 143 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 33213

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 577
- Ceres seq_id 1482457

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 578
- Ceres seq_id 1482458
- Location of start within SEQ ID NO 577: at 1 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 579
- Ceres seq_id 1482459
- Location of start within SEQ ID NO 577: at 35 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 35310

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 580
- Ceres seq_id 1482460

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 581
- Ceres seq_id 1482461
- Location of start within SEQ ID NO 580: at 119 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 582
- Ceres seq_id 1482462
- Location of start within SEQ ID NO 580: at 203 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 583
- Ceres seq_id 1482463
- Location of start within SEQ ID NO 580: at 470 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 37200

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 584
- Ceres seq_id 1482481

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 585
- Ceres seq_id 1482482
- Location of start within SEQ ID NO 584: at 110 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 586
- Ceres seq_id 1482483
- Location of start within SEQ ID NO 584: at 233 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 587
- Ceres seq_id 1482484
- Location of start within SEQ ID NO 584: at 425 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 38293

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 588
- Ceres seq_id 1482490

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 589
- Ceres seq_id 1482491
- Location of start within SEQ ID NO 588: at 104 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 590
- Ceres seq_id 1482492
- Location of start within SEQ ID NO 588: at 138 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 591
- Ceres seq_id 1482493
- Location of start within SEQ ID NO 588: at 151 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 40190

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 592
- Ceres seq_id 1482504

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 593
- Ceres seq_id 1482505
- Location of start within SEQ ID NO 592: at 113 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 594
 - Ceres seq_id 1482506
 - Location of start within SEQ ID NO 592: at 149 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 595
 - Ceres seq_id 1482507
 - Location of start within SEQ ID NO 592: at 642 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 4026

- (A) Polynucleotide Sequence
- Pat. Appln. SEQ ID NO 596
 - Ceres seq_id 1482508
- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 597
 - Ceres seq_id 1482509
 - Location of start within SEQ ID NO 596: at 139 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 598
 - Ceres seq_id 1482510
 - Location of start within SEQ ID NO 596: at 475 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 40770

- (A) Polynucleotide Sequence
- Pat. Appln. SEQ ID NO 599
 - Ceres seq_id 1482514
- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 600
 - Ceres seq_id 1482515
 - Location of start within SEQ ID NO 599: at 33 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)
(D) Related Amino Acid Sequences

- (B) Polypeptide Sequence
- Pat. Appln. SEQ ID NO 601
 - Ceres seq_id 1482516
 - Location of start within SEQ ID NO 599: at 39 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 602
- Ceres seq_id 1482517
- Location of start within SEQ ID NO 599: at 66 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 6091

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 603
- Ceres seq_id 1482525

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 604
- Ceres seq_id 1482526
- Location of start within SEQ ID NO 603: at 79 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 9184

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 605
- Ceres seq_id 1482535

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 606
- Ceres seq_id 1482536
- Location of start within SEQ ID NO 605: at 3 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 607
- Ceres seq_id 1482537
- Location of start within SEQ ID NO 605: at 33 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 608
- Ceres seq_id 1482538
- Location of start within SEQ ID NO 605: at 198 nt.

(C) Nomination and Annotation of Domains within Predicted
Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 92491

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 609
- Ceres seq_id 1482542

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 610
- Ceres seq_id 1482543
- Location of start within SEQ ID NO 609: at 2 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 611
- Ceres seq_id 1482544
- Location of start within SEQ ID NO 609: at 227 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 612
- Ceres seq_id 1482545
- Location of start within SEQ ID NO 609: at 275 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

Maximum Length Sequence corresponding to clone ID 93534

(A) Polynucleotide Sequence

- Pat. Appln. SEQ ID NO 613
- Ceres seq_id 1482546

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 614
- Ceres seq_id 1482547
- Location of start within SEQ ID NO 613: at 218 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 615
- Ceres seq_id 1482548
- Location of start within SEQ ID NO 613: at 227 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

(D) Related Amino Acid Sequences

(B) Polypeptide Sequence

- Pat. Appln. SEQ ID NO 616
- Ceres seq_id 1482549
- Location of start within SEQ ID NO 613: at 260 nt.

(C) Nomination and Annotation of Domains within Predicted Polypeptide(s)

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031	2032	2033	2034	2035	2036	2037	2038	2039	2040	2041	2042	2043	2044	2045	2046	2047	2048	2049	2050	2051	2052	2053	2054	2055	2056	2057	2058	2059	2060	2061	2062	2063	2064	2065	2066	2067	2068	2069	2070	2071	2072	2073	2074	2075	2076	2077	2078	2079	2080	2081	2082	2083	2084	2085	2086	2087	2088	2089	2090	2091	2092	2093	2094	2095	2096	2097	2098	2099	2100	2101	2102	2103	2104	2105	2106	2107	2108	2109	2110	2111	2112	2113	2114	2115	2116	2117	2118	2119	2120	2121	2122	2123	2124	2125	2126	2127	2128	2129	2130	2131	2132	2133	2134	2135	2136	2137	2138	2139	2140	2141	2142	2143	2144	2145	2146	2147	2148	2149	2150	2151	2152	2153	2154	2155	2156	2157	2158	2159	2160	2161	2162	2163	2164	2165	2166	2167	2168	2169	2170	2171	2172	2173	2174	2175	2176	2177	2178	2179	2180	2181	2182	2183	2184	2185	2186	2187	2188	2189	2190	2191	2192	2193	2194	2195	2196	2197	2198	2199	2200	2201	2202	2203	2204	2205	2206	2207	2208	2209	2210	2211	2212	2213	2214	2215	2216	2217	2218	2219	2220	2221	2222	2223	2224	2225	2226	2227	2228	2229	2230	2231	2232	2233	2234	2235	2236	2237	2238	2239	2240	2241	2242	2243	2244	2245	2246	2247	2248	2249	2250	2251	2252	2253	2254	2255	2256	2257	2258	2259	2260	2261	2262	2263	2264	2265	2266	2267	2268	2269	2270	2271	2272	2273	2274	2275	2276	2277	2278	2279	2280	2281	2282	2283	2284	2285	2286	2287	2288	2289	2290	2291	2292	2293	2294	2295	2296	2297	2298	2299	2300	2301	2302	2303	2304	2305	2306	2307	2308	2309	2310	2311	2312	2313	2314	2315	2316	2317	2318	2319	2320	2321	2322	2323	2324	2325	2326	2327	2328	2329	2330	2331	2332	2333	2334	2335	2336	2337	2338	2339	2340	2341	2342	2343	2344	2345	2346	2347	2348	2349	2350	2351	2352	2353	2354	2355	2356	2357	2358	2359	2360	2361	2362	2363	2364	2365	2366	2367	2368	2369	2370	2371	2372	2373	2374	2375	2376	2377	2378	2379	2380	2381	2382	2383	2384	2385	2386	2387	2388	2389	2390	2391	2392	2393	2394	2395	2396	2397	2398	2399	2400	2401	2402	2403	2404	2405	2406	2407
--	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------

Table 1
Page 80

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1458 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1458
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481332

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

```
atccgtttcg ccatttttgt ttctcagtga tctctgaaat ggtctcttct ctttttttgg      60
tcgaatccaa tctcaattat gttgttatct ttcttccatc aatgggtaat caaaacatag      120
aattgatgcg gtaagactat aaaggtttag tctttaacca ttgtagattc ctctgtctct      180
tgtgtatttg attgatctgt taatggataa ccaaaaagggt gctctcttct ccgatgaggt      240
tattctccag attcttgcta gattacctgt taaatctctc ttcaggttca aatccgtttg      300
caaatcatgg tacagattac cttctgacaa atatttctact tccttgttca atcaactctc      360
tgtaaaagag caattgcttg tggctcaagt atcagattct tctagtttga tctgtgttga      420
taatctgaga ggtgtttctg agttatcatt ggattttgtt agagataggg tgaggattag      480
ggtttcttct aatggtttgt tgtgtgttgc aagcattcct gaaaagggtg tttactatgt      540
ttgtaatccg tcgactagag agtacaggaa attgcctaag agtcgagaaa gaccggttac      600
tcgggtttat cctgacgggt aggtacact tgttggtttg gcttgtgatt tgagtaggaa      660
caagttaaat gtggtgttgg ctggttacca taggtctttt ggtcagagac ctgatgggag      720
tttcatttgc ttggtgttgc attctgagag taacaaatgg aggaagtgtt tttcgggtgtt      780
agaagaatgt agtttcacac acatgagtaa gaaccaagtg gtgtttgtta atgggatgct      840
tcattgggtg atgagtgtgt tgtgttatat acttgcactt gatgttgaac atgatgtgtg      900
gagaaagatt tctttgcctg atgagattaa aatcgggaat ggtgggtggta atcgggttta      960
tctcttgtaa tccgatgggt ttttgtcggg gattcagtta tcagatgtat ggatgaagat     1020
ttggaagatg agtgagtatg agactgaaac ttggagtgtt gttgatagca taagttaaag     1080
gtgcattaaa ggatttgtac ctggaatcct cccgatttgt cagaccgggt agtatgtttt     1140
cttggctact cataaacagg ttttgggtga tcaaagacga agtaagttat ggaaagagat     1200
gttttctgta aaaggaagct cttctctgcc tttgtgttgc tctgtctcac cctttcgcag     1260
caccatagta ccctgtaatt agcatgttta tgtttccttc tctactcttt tatttttttg     1320
gtttatgttc agctcttgga tcttttaggg cttatgaaaa tttgttcaag gttttataat     1380
ctttctggga taacatcata taaagtaatg tacagttgat ttcttctgtt gcttttagta     1440
caaataagagt tttggttg
```

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 359 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..359
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481333

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

```
Met Asp Asn Gln Lys Gly Ala Leu Phe Pro Asp Glu Val Ile Leu Gln
1           5           10           15
Ile Leu Ala Arg Leu Pro Val Lys Ser Leu Phe Arg Phe Lys Ser Val
20           25           30
Cys Lys Ser Trp Tyr Arg Leu Pro Ser Asp Lys Tyr Phe Thr Ser Leu
35           40           45
Phe Asn Gln Leu Ser Val Lys Glu Gln Leu Leu Val Ala Gln Val Ser
50           55           60
Asp Ser Ser Ser Leu Ile Cys Val Asp Asn Leu Arg Gly Val Ser Glu
65           70           75           80
```

Leu Ser Leu Asp Phe Val Arg Asp Arg Val Arg Ile Arg Val Ser Ser
85 90 95
Asn Gly Leu Leu Cys Cys Ser Ser Ile Pro Glu Lys Gly Val Tyr Tyr
100 105 110
Val Cys Asn Pro Ser Thr Arg Glu Tyr Arg Lys Leu Pro Lys Ser Arg
115 120 125
Glu Arg Pro Val Thr Arg Phe Tyr Pro Asp Gly Glu Ala Thr Leu Val
130 135 140
Gly Leu Ala Cys Asp Leu Ser Arg Asn Lys Phe Asn Val Val Leu Ala
145 150 155 160
Gly Tyr His Arg Ser Phe Gly Gln Arg Pro Asp Gly Ser Phe Ile Cys
165 170 175
Leu Val Phe Asp Ser Glu Ser Asn Lys Trp Arg Lys Phe Val Ser Val
180 185 190
Leu Glu Glu Cys Ser Phe Thr His Met Ser Lys Asn Gln Val Val Phe
195 200 205
Val Asn Gly Met Leu His Trp Leu Met Ser Gly Leu Cys Tyr Ile Leu
210 215 220
Ala Leu Asp Val Glu His Asp Val Trp Arg Lys Ile Ser Leu Pro Asp
225 230 235 240
Glu Ile Lys Ile Gly Asn Gly Gly Gly Asn Arg Val Tyr Leu Leu Glu
245 250 255
Ser Asp Gly Phe Leu Ser Val Ile Gln Leu Ser Asp Val Trp Met Lys
260 265 270
Ile Trp Lys Met Ser Glu Tyr Glu Thr Glu Thr Trp Ser Val Val Asp
275 280 285
Ser Ile Ser Leu Arg Cys Ile Lys Gly Leu Val Pro Gly Ile Phe Pro
290 295 300
Ile Cys Gln Thr Gly Glu Tyr Val Phe Leu Ala Thr His Lys Gln Val
305 310 315 320
Leu Val Tyr Gln Arg Ser Lys Leu Trp Lys Glu Met Phe Ser Val
325 330 335
Lys Gly Ser Ser Ser Leu Pro Leu Trp Phe Ser Ala His Ala Phe Arg
340 345 350
Ser Thr Ile Val Pro Cys Asn
355

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1353 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1353
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481342

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

casragamcc atwacywaga amcaycctaa tcgaaaaaac gccacaatca tggctttggt	60
cttatctcct aaaaccatca ctcttctctt cttctccctc tccctcgcac tctactgcag	120
catcgatcct ttccaccact gcgccatttc cgatttcccc aatttcgtct ctcacgaagt	180
tatctctcca cgtcccgcagc aagttccatg ggagagagat tcacaaaatt cacttcagaa	240
atcaaagatt ctgtttttta accaaatcca aggtccagag agcgtcgcct ttgattctct	300
cggacgtggt ccgtacacag gcgttgctga tggtaggggt ttgttttggg atggagagaa	360
atggattgat ttgcgttata cttcgagtaa tcgatcggag atttgtgatc cgaagccttc	420
tgctttgagt tacttgagga atgaacatat atgtggtcgt cctttagggtc ttcgtttcga	480
taagagaacc ggagatttgt atatagctga tgcttatatg ggacttttga aagttgggtcc	540
tgaaggtggt ttagcaacgc cgcttgtaac tgaagctgaa ggtgtgccgt tgggggtttac	600
taatgatctt gacattgctg atgatggaac tgtttacttt acagatagca gcattagtta	660

```
ccagaggagg aacttcttgc agctcgtttt ctctggagac aatactggga gggttctaaa 720
gtatgatcca gtagctaaga aagctgttgt tttggtctca aatcttcagt ttccgaatgg 780
tgtctctatc agcagagacg gttcttttct tgtattctgc gaaggagata ttggaagcct 840
acgaagatac tgggttgaaag gcgagaaagc tggaacgaca gatgtgtttg cgtattttacc 900
agggcatcct gataacgtaa gaaccaacca aaagggtgaa ttttgggtag cgcttcattg 960
cagacgcaac tactactcat acttaatggc aagatatcct aagctgagga tgttcatact 1020
gagactgcca atcactgcga gaactcacta ctctgtccag ataggggttac ggccgcacgg 1080
gttggtggtt aagtatagtc ctgaagggaa gcttatgcat gttttggaag atagtgaagg 1140
gaaagttgtg agatcagtaa gtgaagtgga agaaaaagat gggaagcttt ggatgggaag 1200
tgtgttgatg aactttgttg ctgtctatga cctctgatta cttgacctat acgtaaacca 1260
cttcactcag tttctagatt tagcaaattc tcaaaactgt taggtgtgta ctgaaaaaat 1320
caaacactta gcacaaacaa actcaatggt att
```

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 411 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..411

(D) OTHER INFORMATION: / Ceres Seq. ID 1481343

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

```
Xaa Arg Xaa Xaa Xaa Xaa Xaa Pro Asn Arg Lys Asn Ala Thr Ile
1      5      10      15
Met Ala Leu Phe Leu Ser Pro Lys Thr Ile Thr Leu Leu Phe Phe Ser
20      25      30
Leu Ser Leu Ala Leu Tyr Cys Ser Ile Asp Pro Phe His His Cys Ala
35      40      45
Ile Ser Asp Phe Pro Asn Phe Val Ser His Glu Val Ile Ser Pro Arg
50      55      60
Pro Asp Glu Val Pro Trp Glu Arg Asp Ser Gln Asn Ser Leu Gln Lys
65      70      75      80
Ser Lys Ile Leu Phe Phe Asn Gln Ile Gln Gly Pro Glu Ser Val Ala
85      90      95
Phe Asp Ser Leu Gly Arg Gly Pro Tyr Thr Gly Val Ala Asp Gly Arg
100      105      110
Val Leu Phe Trp Asp Gly Glu Lys Trp Ile Asp Phe Ala Tyr Thr Ser
115      120      125
Ser Asn Arg Ser Glu Ile Cys Asp Pro Lys Pro Ser Ala Leu Ser Tyr
130      135      140
Leu Arg Asn Glu His Ile Cys Gly Arg Pro Leu Gly Leu Arg Phe Asp
145      150      155      160
Lys Arg Thr Gly Asp Leu Tyr Ile Ala Asp Ala Tyr Met Gly Leu Leu
165      170      175
Lys Val Gly Pro Glu Gly Gly Leu Ala Thr Pro Leu Val Thr Glu Ala
180      185      190
Glu Gly Val Pro Leu Gly Phe Thr Asn Asp Leu Asp Ile Ala Asp Asp
195      200      205
Gly Thr Val Tyr Phe Thr Asp Ser Ser Ile Ser Tyr Gln Arg Arg Asn
210      215      220
Phe Leu Gln Leu Val Phe Ser Gly Asp Asn Thr Gly Arg Val Leu Lys
225      230      235      240
Tyr Asp Pro Val Ala Lys Lys Ala Val Val Leu Val Ser Asn Leu Gln
245      250      255
Phe Pro Asn Gly Val Ser Ile Ser Arg Asp Gly Ser Phe Phe Val Phe
260      265      270
Cys Glu Gly Asp Ile Gly Ser Leu Arg Arg Tyr Trp Leu Lys Gly Glu
```

275	280	285
Lys Ala Gly Thr Thr Asp Val Phe Ala Tyr Leu Pro Gly His Pro Asp		
290	295	300
Asn Val Arg Thr Asn Gln Lys Gly Glu Phe Trp Val Ala Leu His Cys		
305	310	315
Arg Arg Asn Tyr Tyr Ser Tyr Leu Met Ala Arg Tyr Pro Lys Leu Arg		
	325	330
Met Phe Ile Leu Arg Leu Pro Ile Thr Ala Arg Thr His Tyr Ser Phe		
	340	345
Gln Ile Gly Leu Arg Pro His Gly Leu Val Val Lys Tyr Ser Pro Glu		
	355	360
Gly Lys Leu Met His Val Leu Glu Asp Ser Glu Gly Lys Val Val Arg		
	370	375
Ser Val Ser Glu Val Glu Lys Asp Gly Lys Leu Trp Met Gly Ser		
385	390	395
Val Leu Met Asn Phe Val Ala Val Tyr Asp Leu		
	405	410

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 395 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..395

(D) OTHER INFORMATION: / Ceres Seq. ID 1481344

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

Met Ala Leu Phe Leu Ser Pro Lys Thr Ile Thr Leu Leu Phe Phe Ser		
1	5	10
Leu Ser Leu Ala Leu Tyr Cys Ser Ile Asp Pro Phe His His Cys Ala		
	20	25
Ile Ser Asp Phe Pro Asn Phe Val Ser His Glu Val Ile Ser Pro Arg		
	35	40
Pro Asp Glu Val Pro Trp Glu Arg Asp Ser Gln Asn Ser Leu Gln Lys		
	50	55
Ser Lys Ile Leu Phe Phe Asn Gln Ile Gln Gly Pro Glu Ser Val Ala		
65	70	75
Phe Asp Ser Leu Gly Arg Gly Pro Tyr Thr Gly Val Ala Asp Gly Arg		
	85	90
Val Leu Phe Trp Asp Gly Glu Lys Trp Ile Asp Phe Ala Tyr Thr Ser		
	100	105
Ser Asn Arg Ser Glu Ile Cys Asp Pro Lys Pro Ser Ala Leu Ser Tyr		
	115	120
Leu Arg Asn Glu His Ile Cys Gly Arg Pro Leu Gly Leu Arg Phe Asp		
	130	135
Lys Arg Thr Gly Asp Leu Tyr Ile Ala Asp Ala Tyr Met Gly Leu Leu		
145	150	155
Lys Val Gly Pro Glu Gly Gly Leu Ala Thr Pro Leu Val Thr Glu Ala		
	165	170
Glu Gly Val Pro Leu Gly Phe Thr Asn Asp Leu Asp Ile Ala Asp Asp		
	180	185
Gly Thr Val Tyr Phe Thr Asp Ser Ser Ile Ser Tyr Gln Arg Arg Asn		
	195	200
Phe Leu Gln Leu Val Phe Ser Gly Asp Asn Thr Gly Arg Val Leu Lys		
	210	215
Tyr Asp Pro Val Ala Lys Lys Ala Val Val Leu Val Ser Asn Leu Gln		
225	230	235
		240

Phe Pro Asn Gly Val Ser Ile Ser Arg Asp Gly Ser Phe Phe Val Phe
245 250 255
Cys Glu Gly Asp Ile Gly Ser Leu Arg Arg Tyr Trp Leu Lys Gly Glu
260 265 270
Lys Ala Gly Thr Thr Asp Val Phe Ala Tyr Leu Pro Gly His Pro Asp
275 280 285
Asn Val Arg Thr Asn Gln Lys Gly Glu Phe Trp Val Ala Leu His Cys
290 295 300
Arg Arg Asn Tyr Tyr Ser Tyr Leu Met Ala Arg Tyr Pro Lys Leu Arg
305 310 315 320
Met Phe Ile Leu Arg Leu Pro Ile Thr Ala Arg Thr His Tyr Ser Phe
325 330 335
Gln Ile Gly Leu Arg Pro His Gly Leu Val Val Lys Tyr Ser Pro Glu
340 345 350
Gly Lys Leu Met His Val Leu Glu Asp Ser Glu Gly Lys Val Val Arg
355 360 365
Ser Val Ser Glu Val Glu Glu Lys Asp Gly Lys Leu Trp Met Gly Ser
370 375 380
Val Leu Met Asn Phe Val Ala Val Tyr Asp Leu
385 390 395

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 239 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..239

(D) OTHER INFORMATION: / Ceres Seq. ID 1481345

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

Met Gly Leu Leu Lys Val Gly Pro Glu Gly Gly Leu Ala Thr Pro Leu
1 5 10 15
Val Thr Glu Ala Glu Gly Val Pro Leu Gly Phe Thr Asn Asp Leu Asp
20 25 30
Ile Ala Asp Asp Gly Thr Val Tyr Phe Thr Asp Ser Ser Ile Ser Tyr
35 40 45
Gln Arg Arg Asn Phe Leu Gln Leu Val Phe Ser Gly Asp Asn Thr Gly
50 55 60
Arg Val Leu Lys Tyr Asp Pro Val Ala Lys Lys Ala Val Val Leu Val
65 70 75 80
Ser Asn Leu Gln Phe Pro Asn Gly Val Ser Ile Ser Arg Asp Gly Ser
85 90 95
Phe Phe Val Phe Cys Glu Gly Asp Ile Gly Ser Leu Arg Arg Tyr Trp
100 105 110
Leu Lys Gly Glu Lys Ala Gly Thr Thr Asp Val Phe Ala Tyr Leu Pro
115 120 125
Gly His Pro Asp Asn Val Arg Thr Asn Gln Lys Gly Glu Phe Trp Val
130 135 140
Ala Leu His Cys Arg Arg Asn Tyr Tyr Ser Tyr Leu Met Ala Arg Tyr
145 150 155 160
Pro Lys Leu Arg Met Phe Ile Leu Arg Leu Pro Ile Thr Ala Arg Thr
165 170 175
His Tyr Ser Phe Gln Ile Gly Leu Arg Pro His Gly Leu Val Val Lys
180 185 190
Tyr Ser Pro Glu Gly Lys Leu Met His Val Leu Glu Asp Ser Glu Gly
195 200 205
Lys Val Val Arg Ser Val Ser Glu Val Glu Glu Lys Asp Gly Lys Leu

210 215 220
Trp Met Gly Ser Val Leu Met Asn Phe Val Ala Val Tyr Asp Leu
225 230 235

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1279 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1279
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481346

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

```
atcaccattg ctttgttttg ttcgtaaaat ataatcaatt tttaaatttct tctcttctct 60
tcaaacgaat cgcctttttc gataatctct ttgcatcgat ttcatcatgg ctactcaaac 120
ggatctcgct cagcccaagc ttgatatgac caaggaggag aaagagaggt tgaagtattt 180
gcaattcgct caagctgctg ctgtggaagc tctgcttcgc tttgctctta ttacgctaa 240
ggcaaaggac aagctctggtc ctttgaaacc tgggtgttgaa tctgttgaaag gagctgtcaa 300
gactgtcggt ggtcctgtct acgagaaata ccacgacgct cctgttgagg tccttaaata 360
catggaccag aaggtacaat tttgactott tccctatctt tggatcttgc tgaaagtgcc 420
tttgttgatg aacaatgaat gaatctgtgt tgttgattgt atatccactt catcgaacat 480
atgtgattaa aaaagtacag ttaaagtgtg gatgatttca tatcatctct ttggtagaag 540
gttcaggtta acgggtcaat gtcattatgt tctgtagagt ccctcttttt gaagctgaca 600
agtttgtttt gcgttggtgc aggttgatat gtctgtgact gagcttgacc gtcgtgtccc 660
accagtcgtc aagcaagtgt ctgcccagc catctccgct gctcagatag caccattgt 720
ggcacgtgcg ttggcctctg aggttcgacg tgctgggtgt gttgaaaccg cttctggaat 780
ggctaaatcc gtctactcca agtacgagcc tgctgctaag gagttgtatg caaactatga 840
gccaaaagca aagcagtgtg ccgtttcagc ttggaagaag ctttaaccagc ttctcttatt 900
cccaaggctg gctcaagttg ctgtaccaac agctgctttc tgctctgaga agtacaatga 960
tactgtggtt aaggctgcag agaaagggtg cagagtcaca tegtacatgc cattgggttc 1020
aacagagagg atctcaaaaa tcttcgctga ggagaaagct gagaccgagc ctttgaggtt 1080
ccatccactt gattgatatg ggtgttttgt tagtgtgatt ttttgttttg ttgggattaa 1140
ggtgaaccgg atcttggtta gcgattgatc tctggttctc gttctttttt ttctttgtca 1200
tgaacttttg ttgtttcggt taataatcaa aagttgtata atctaagttt gggattacca 1260
ccctattgag tattgagtg
```

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 155 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..155
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481347

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

```
Met Ser Val Thr Glu Leu Asp Arg Arg Val Pro Pro Val Val Lys Gln
1 5 10 15
Val Ser Ala Gln Ala Ile Ser Ala Ala Gln Ile Ala Pro Ile Val Ala
20 25 30
Arg Ala Leu Ala Ser Glu Val Arg Arg Ala Gly Val Val Glu Thr Ala
35 40 45
Ser Gly Met Ala Lys Ser Val Tyr Ser Lys Tyr Glu Pro Ala Ala Lys
50 55 60
Glu Leu Tyr Ala Asn Tyr Glu Pro Lys Ala Lys Gln Cys Ala Val Ser
65 70 75 80
```

Ala Trp Lys Lys Leu Asn Gln Leu Pro Leu Phe Pro Arg Leu Ala Gln
85 90 95
Val Ala Val Pro Thr Ala Ala Phe Cys Ser Glu Lys Tyr Asn Asp Thr
100 105 110
Val Val Lys Ala Ala Glu Lys Gly Tyr Arg Val Thr Ser Tyr Met Pro
115 120 125
Leu Val Pro Thr Glu Arg Ile Ser Lys Ile Phe Ala Glu Glu Lys Ala
130 135 140
Glu Thr Glu Pro Leu Glu Phe His Pro Leu Asp
145 150 155

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 105 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..105
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481348

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

Met Ala Lys Ser Val Tyr Ser Lys Tyr Glu Pro Ala Ala Lys Glu Leu
1 5 10 15
Tyr Ala Asn Tyr Glu Pro Lys Ala Lys Gln Cys Ala Val Ser Ala Trp
20 25 30
Lys Lys Leu Asn Gln Leu Pro Leu Phe Pro Arg Leu Ala Gln Val Ala
35 40 45
Val Pro Thr Ala Ala Phe Cys Ser Glu Lys Tyr Asn Asp Thr Val Val
50 55 60
Lys Ala Ala Glu Lys Gly Tyr Arg Val Thr Ser Tyr Met Pro Leu Val
65 70 75 80
Pro Thr Glu Arg Ile Ser Lys Ile Phe Ala Glu Glu Lys Ala Glu Thr
85 90 95
Glu Pro Leu Glu Phe His Pro Leu Asp
100 105

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 96 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..96
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481349

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

Met Gln Thr Met Ser Gln Lys Gln Ser Ser Val Pro Phe Gln Leu Gly
1 5 10 15
Arg Ser Leu Thr Ser Phe Leu Tyr Ser Gln Gly Trp Leu Lys Leu Leu
20 25 30
Tyr Gln Gln Leu Leu Ser Ala Leu Arg Ser Thr Met Ile Leu Trp Leu
35 40 45
Arg Leu Gln Arg Lys Gly Thr Glu Ser His Arg Thr Cys His Trp Phe
50 55 60
Gln Gln Arg Gly Ser Gln Lys Ser Ser Leu Arg Arg Lys Leu Arg Pro
65 70 75 80
Ser Leu Trp Ser Ser Ile His Leu Ile Asp Met Gly Val Leu Leu Val

85

90

95

(2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1211 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1211
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481357

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

```
aacatcctaa tcgaaaaaaa aaaacataaa acacataggg gtgggtctct ctccctccgga      60
attcgatcac gacggcaagg acgacgcac tccttctccc acagggctgg agatggatct      120
gggtccggtga tttctgagat ttaagtcgat cgagtttcca gatatatctc tcaagtagag      180
atggcttggt tcagtggcaa agtttctctg ggaggattcc cagatctcac tggcgctgtc      240
aataaattcc agagagcggt aaaaacattg aaaagaattt cgacaacgcc cttgggttcg      300
acgacaagtc cgattctgcc gctgaagatg cagcttcaag tatgtggcca cctgcagttg      360
ataccaaaag cctctttgat cccgttatgt ccttcatggg taacacctct gatgagaaac      420
ctgatacatt ggaagactct gtgcgtacag aaaatccgtc tcaaattgaa caaaaagaag      480
aagaagctgg atcggttaag ctagtactg aacaagcagt atctgttgaa gcaaataaag      540
aaacaaacat gagaagagaa gctgatcaag cagataatcc tgaggtaaca gaaactgttg      600
ttttggatcc caacgatgat gaaccgcaat cgcagatact tctcgaagag tcctctgaat      660
attctcttca gactcctgaa tcctcaggtt acaagactag tcttcaacct aatgaaaagc      720
tggaatgac agcttctcaa gattcacagc ccgagcaacc caagtcagag gctgaggaat      780
cacagcctga ggattctgaa gcaaaagagg ttactgtaga aaacaaagac actgttcact      840
cccctgtgtt agatggacag cataagatta cttatatgga tgagacaaca aatgaacaag      900
aaattctggg tgaatatctg gaagggagaa cctcgtctaa aatttttgaa gtttcaccag      960
atatcaatca tgtaaataag atagagtccc ttgttgctca tccgtcttta atttttgagt      1020
ctgatgggtt tccttacgag tcttctatac caaagagatc gtcgtcagat gaaatttcgg      1080
agagaattgt ggattttgtt tctcgtgaaa tagattcaag actggatact agtgagttaa      1140
atgaaagcca gcgttcaagc tctgcgacaa atgtttccga ctctgctgat gttattctgg      1200
aattagagaa g
```

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 290 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..290
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481358

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

```
Met Trp Pro Pro Ala Val Asp Thr Lys Ser Leu Phe Asp Pro Val Met
1          5          10          15
Ser Phe Met Gly Asn Thr Ser Asp Glu Lys Pro Asp Thr Leu Glu Asp
20          25          30
Ser Val Arg Thr Glu Asn Pro Ser Gln Ile Glu Gln Lys Glu Glu Glu
35          40          45
Ala Gly Ser Val Lys Leu Ala Thr Glu Gln Ala Val Ser Val Glu Ala
50          55          60
Asn Lys Glu Thr Asn Met Arg Arg Glu Ala Asp Gln Ala Asp Asn Pro
65          70          75          80
Glu Val Thr Glu Thr Val Val Leu Asp Pro Asn Asp Asp Glu Pro Gln
```

				85					90					95			
Ser	Gln	Ile	Leu	Leu	Glu	Glu	Ser	Ser	Glu	Tyr	Ser	Leu	Gln	Thr	Pro		
			100					105					110				
Glu	Ser	Ser	Gly	Tyr	Lys	Thr	Ser	Leu	Gln	Pro	Asn	Glu	Lys	Leu	Glu		
		115					120					125					
Met	Thr	Ala	Ser	Gln	Asp	Ser	Gln	Pro	Glu	Gln	Pro	Lys	Ser	Glu	Ala		
	130					135					140						
Glu	Glu	Ser	Gln	Pro	Glu	Asp	Ser	Glu	Ala	Lys	Glu	Val	Thr	Val	Glu		
	145				150					155					160		
Asn	Lys	Asp	Thr	Val	His	Ser	Pro	Val	Leu	Asp	Gly	Gln	His	Lys	Ile		
			165					170						175			
Thr	Tyr	Met	Asp	Glu	Thr	Thr	Asn	Glu	Gln	Glu	Ile	Leu	Gly	Glu	Asn		
		180					185					190					
Leu	Glu	Gly	Arg	Thr	Ser	Ser	Lys	Ile	Phe	Glu	Val	Ser	Pro	Asp	Ile		
	195						200					205					
Asn	His	Val	Asn	Arg	Ile	Glu	Ser	Leu	Val	Ala	His	Pro	Ser	Leu	Ile		
	210				215						220						
Phe	Glu	Ser	Asp	Gly	Ser	Pro	Tyr	Glu	Ser	Ser	Ile	Pro	Lys	Arg	Ser		
	225			230						235					240		
Ser	Ser	Asp	Glu	Ile	Ser	Glu	Arg	Ile	Val	Asp	Phe	Val	Ser	Arg	Glu		
			245					250						255			
Ile	Asp	Ser	Arg	Leu	Asp	Thr	Ser	Glu	Leu	Asn	Glu	Ser	Gln	Arg	Ser		
		260						265					270				
Ser	Ser	Ala	Thr	Asn	Val	Ser	Asp	Ser	Ala	Asp	Val	Ile	Leu	Glu	Leu		
		275					280					285					
Glu	Lys																
	290																

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 275 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..275
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481359

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

Met	Ser	Phe	Met	Gly	Asn	Thr	Ser	Asp	Glu	Lys	Pro	Asp	Thr	Leu	Glu		
1			5					10						15			
Asp	Ser	Val	Arg	Thr	Glu	Asn	Pro	Ser	Gln	Ile	Glu	Gln	Lys	Glu	Glu		
		20					25						30				
Glu	Ala	Gly	Ser	Val	Lys	Leu	Ala	Thr	Glu	Gln	Ala	Val	Ser	Val	Glu		
	35					40					45						
Ala	Asn	Lys	Glu	Thr	Asn	Met	Arg	Arg	Glu	Ala	Asp	Gln	Ala	Asp	Asn		
	50				55				60								
Pro	Glu	Val	Thr	Glu	Thr	Val	Val	Leu	Asp	Pro	Asn	Asp	Asp	Glu	Pro		
	65			70				75						80			
Gln	Ser	Gln	Ile	Leu	Glu	Glu	Ser	Ser	Glu	Tyr	Ser	Leu	Gln	Thr			
		85					90					95					
Pro	Glu	Ser	Ser	Gly	Tyr	Lys	Thr	Ser	Leu	Gln	Pro	Asn	Glu	Lys	Leu		
	100						105					110					
Glu	Met	Thr	Ala	Ser	Gln	Asp	Ser	Gln	Pro	Glu	Gln	Pro	Lys	Ser	Glu		
	115					120						125					
Ala	Glu	Glu	Ser	Gln	Pro	Glu	Asp	Ser	Glu	Ala	Lys	Glu	Val	Thr	Val		
	130				135					140							
Glu	Asn	Lys	Asp	Thr	Val	His	Ser	Pro	Val	Leu	Asp	Gly	Gln	His	Lys		
	145				150				155						160		

Ile	Thr	Tyr	Met	Asp	Glu	Thr	Thr	Asn	Glu	Gln	Glu	Ile	Leu	Gly	Glu
			165						170					175	
Asn	Leu	Glu	Gly	Arg	Thr	Ser	Ser	Lys	Ile	Phe	Glu	Val	Ser	Pro	Asp
			180					185					190		
Ile	Asn	His	Val	Asn	Arg	Ile	Glu	Ser	Leu	Val	Ala	His	Pro	Ser	Leu
		195					200					205			
Ile	Phe	Glu	Ser	Asp	Gly	Ser	Pro	Tyr	Glu	Ser	Ser	Ile	Pro	Lys	Arg
	210					215					220				
Ser	Ser	Ser	Asp	Glu	Ile	Ser	Glu	Arg	Ile	Val	Asp	Phe	Val	Ser	Arg
225					230					235					240
Glu	Ile	Asp	Ser	Arg	Leu	Asp	Thr	Ser	Glu	Leu	Asn	Glu	Ser	Gln	Arg
				245					250					255	
Ser	Ser	Ser	Ala	Thr	Asn	Val	Ser	Asp	Ser	Ala	Asp	Val	Ile	Leu	Glu
			260					265					270		
Leu	Glu	Lys													
		275													

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 272 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..272

(D) OTHER INFORMATION: / Ceres Seq. ID 1481360

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

Met	Gly	Asn	Thr	Ser	Asp	Glu	Lys	Pro	Asp	Thr	Leu	Glu	Asp	Ser	Val
1			5					10						15	
Arg	Thr	Glu	Asn	Pro	Ser	Gln	Ile	Glu	Gln	Lys	Glu	Glu	Glu	Ala	Gly
		20					25					30			
Ser	Val	Lys	Leu	Ala	Thr	Glu	Gln	Ala	Val	Ser	Val	Glu	Ala	Asn	Lys
		35				40					45				
Glu	Thr	Asn	Met	Arg	Arg	Glu	Ala	Asp	Gln	Ala	Asp	Asn	Pro	Glu	Val
	50					55				60					
Thr	Glu	Thr	Val	Val	Leu	Asp	Pro	Asn	Asp	Asp	Glu	Pro	Gln	Ser	Gln
65					70				75					80	
Ile	Leu	Leu	Glu	Glu	Ser	Ser	Glu	Tyr	Ser	Leu	Gln	Thr	Pro	Glu	Ser
			85				90						95		
Ser	Gly	Tyr	Lys	Thr	Ser	Leu	Gln	Pro	Asn	Glu	Lys	Leu	Glu	Met	Thr
		100					105					110			
Ala	Ser	Gln	Asp	Ser	Gln	Pro	Glu	Gln	Pro	Lys	Ser	Glu	Ala	Glu	Glu
		115					120					125			
Ser	Gln	Pro	Glu	Asp	Ser	Glu	Ala	Lys	Glu	Val	Thr	Val	Glu	Asn	Lys
		130				135					140				
Asp	Thr	Val	His	Ser	Pro	Val	Leu	Asp	Gly	Gln	His	Lys	Ile	Thr	Tyr
145					150				155					160	
Met	Asp	Glu	Thr	Thr	Asn	Glu	Gln	Glu	Ile	Leu	Gly	Glu	Asn	Leu	Glu
			165					170					175		
Gly	Arg	Thr	Ser	Ser	Lys	Ile	Phe	Glu	Val	Ser	Pro	Asp	Ile	Asn	His
			180				185						190		
Val	Asn	Arg	Ile	Glu	Ser	Leu	Val	Ala	His	Pro	Ser	Leu	Ile	Phe	Glu
		195					200					205			
Ser	Asp	Gly	Ser	Pro	Tyr	Glu	Ser	Ser	Ile	Pro	Lys	Arg	Ser	Ser	Ser
	210					215					220				
Asp	Glu	Ile	Ser	Glu	Arg	Ile	Val	Asp	Phe	Val	Ser	Arg	Glu	Ile	Asp
225					230					235				240	
Ser	Arg	Leu	Asp	Thr	Ser	Glu	Leu	Asn	Glu	Ser	Gln	Arg	Ser	Ser	Ser

	245		250		255
Ala Thr Asn Val Ser Asp Ser Ala Asp Val Ile Leu Glu Leu Glu Lys					
	260		265		270

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 592 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..592
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481372

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

ctaatacgaaa	aaaaagcgag	aaagaaagac	gaactgatca	gcaatgggaa	gcttaagggt	60
gagcacagtt	gttattgcag	tagtggcctt	tctctccatc	ctcctcatat	ctcctacaga	120
agtagatggg	cgtttagtgt	gtgacactcc	agcgggtaca	tgtacctoga	gctctacttg	180
caatgaccaa	tgcaatacat	ggggcgga	ttatagtga	ggcgaatgtg	cagattcaag	240
ctttcctggt	ttaagtatat	gttattgctg	ccattatgta	gggagcagtg	ctgaaatgga	300
aagcatgtga	ttgcagatga	tagaaaacga	cgtcgctttg	tgtgcgtatg	tgtgtgtttt	360
ttgctaatacg	catgtttatg	ctttcatttc	acatcctatg	ttttgagtg	ttgcctttgt	420
actttgttgt	tgtgcttctg	tttggtttgc	gttgtcaagt	atcaaataaa	gttgaggagt	480
gtttttaaca	aatgattttt	ttattattct	tgtgtatatt	agctaattta	ttttatttaa	540
gagtggttta	tttttatcaa	taataataat	cataattgcg	gtttgttgtg	cg	

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 47 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..47
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481373

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

Leu	Ile	Glu	Lys	Lys	Ala	Arg	Lys	Lys	Asp	Glu	Leu	Ile	Ser	Asn	Gly
1			5				10						15		
Lys	Leu	Lys	Gly	Glu	His	Ser	Cys	Tyr	Cys	Ser	Ser	Gly	Leu	Ser	Leu
			20				25					30			
His	Pro	Pro	His	Ile	Ser	Tyr	Arg	Ser	Arg	Trp	Ala	Phe	Ser	Val	
			35				40					45			

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 88 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..88
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481374

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

Met	Gly	Ser	Leu	Arg	Val	Ser	Thr	Val	Ile	Ala	Val	Val	Ala	Cys
1			5				10					15		

(2) INFORMATION FOR SEQ ID NO:18:

(A) LENGTH: 46 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..46

(D) OTHER INFORMATION: / Ceres Seq. ID 1481375

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

Met	Cys	Val	Phe	Phe	Ala	Asn	Arg	Met	Phe	Met	Leu	Ser	Phe	His	Ile
1				5					10					15	
Leu	Cys	Phe	Glu	Cys	Leu	Pro	Leu	Tyr	Phe	Val	Val	Val	Leu	Leu	Phe
			20					25					30		
Val	Leu	Arg	Cys	Gln	Val	Ser	Asn	Lys	Val	Gly	Val	Cys	Phe		
		35					40					45			

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 1135 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..1135

(D) OTHER INFORMATION: / Ceres Seq. ID 1481388

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

taaatgagat	gaatagaggt	ccacgagcta	agggtttcaa	cagccaagat	ggttccaagg	60
tgatggctgt	gtctttgaag	gagcagagag	tgactgagac	tgagaaactc	agtgaagatg	120
tgtctctttt	agatcccaag	gactacaata	agatagattt	ccctgagacc	tacacagaag	180
caaagtttta	tgtaatcaaa	tcgtacagtg	aagatgatat	tcataaaaagt	atcaaataca	240
gtgtttggtc	cagcactcct	aatggtaaca	agaagctgga	tgccctcatat	aacgaggcaa	300
aacagaagtc	agatggctgt	cccgtgtttc	tactttttctc	tgtaaacact	agtggacaat	360
ttgttggttt	agccgagatg	gtaggccctg	ttgatttcaa	taagactgtt	gaatactggc	420
aacaggacaa	atggatttgt	tgtctccctg	ttaaagtggca	tttctgtaaa	gataatcccta	480
atagctcctt	gaggcatata	actctggaga	acaatgagaa	caagccgggt	actaatagca	540
gagacacaca	ggaagtaaa	ctcgagcaag	gcattaaagt	catcaagatt	ttcaaggacc	600
acgcaagcaa	gacatgcata	ctcgatgatt	ttgagttcta	tgagaatcgt	caaaagatta	660
tccaagaaa	gaaaagcaaa	cacctgcaga	tcaaaaaaca	gacattgggtg	gccaatgcag	720
acaaaggtgt	aatgtcaaaa	attaatcttg	tgaaacctca	agagtctact	acagcctcag	780
aagatgcagc	agcactagga	gttgcggtg	aagtgactaa	agaatcgaaa	gtggtgaaag	840
agaccgagtt	acctgtggag	aaaaatgctg	ttgctactgc	ctgctgaacc	aaccttttgt	900
tttaagtggg	aactgagtg	gctgttttag	gctattttaga	gcgtttctct	agttttgttt	960
ccattcctga	atttgagcac	tttttttttt	tttttttttg	aaccgagttg	agagggtagt	1020
ggcttagtag	atgaagtttt	ggcatgagca	ttcatcatct	tgcagttatt	ctctatccct	1080
ttagtaaatgg	tccaacatat	gaqqatatgg	gtaaaagatt	ggtattgaat	cagct	

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 294 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..294
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481389

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

Asn	Glu	Met	Asn	Arg	Gly	Pro	Arg	Ala	Lys	Gly	Phe	Asn	Ser	Gln	Asp
1			5						10					15	
Gly	Ser	Lys	Val	Met	Ala	Val	Ser	Leu	Lys	Glu	Gln	Arg	Val	Thr	Glu
			20					25					30		
Thr	Glu	Lys	Leu	Ser	Glu	Asp	Val	Ser	Leu	Leu	Asp	Pro	Lys	Asp	Tyr
			35				40					45			
Asn	Lys	Ile	Asp	Phe	Pro	Glu	Thr	Tyr	Thr	Glu	Ala	Lys	Phe	Tyr	Val
			50				55				60				
Ile	Lys	Ser	Tyr	Ser	Glu	Asp	Asp	Ile	His	Lys	Ser	Ile	Lys	Tyr	Ser
65						70				75				80	
Val	Trp	Ser	Ser	Thr	Pro	Asn	Gly	Asn	Lys	Lys	Leu	Asp	Ala	Ser	Tyr
				85				90						95	
Asn	Glu	Ala	Lys	Gln	Lys	Ser	Asp	Gly	Cys	Pro	Val	Phe	Leu	Leu	Phe
			100					105					110		
Ser	Val	Asn	Thr	Ser	Gly	Gln	Phe	Val	Gly	Leu	Ala	Glu	Met	Val	Gly
			115				120					125			
Pro	Val	Asp	Phe	Asn	Lys	Thr	Val	Glu	Tyr	Trp	Gln	Gln	Asp	Lys	Trp
						135					140				
Ile	Gly	Cys	Phe	Pro	Val	Lys	Trp	His	Phe	Val	Lys	Asp	Ile	Pro	Asn
145						150				155				160	
Ser	Ser	Leu	Arg	His	Ile	Thr	Leu	Glu	Asn	Asn	Glu	Asn	Lys	Pro	Val
				165				170						175	
Thr	Asn	Ser	Arg	Asp	Thr	Gln	Glu	Val	Lys	Leu	Glu	Gln	Gly	Ile	Lys
				180				185					190		
Val	Ile	Lys	Ile	Phe	Lys	Asp	His	Ala	Ser	Lys	Thr	Cys	Ile	Leu	Asp
				195			200					205			
Asp	Phe	Glu	Phe	Tyr	Glu	Asn	Arg	Gln	Lys	Ile	Ile	Gln	Glu	Arg	Lys
						215					220				
Ser	Lys	His	Leu	Gln	Ile	Lys	Lys	Gln	Thr	Leu	Val	Ala	Asn	Ala	Asp
225					230					235					240
Lys	Gly	Val	Met	Ser	Lys	Ile	Asn	Leu	Val	Lys	Pro	Gln	Glu	Ser	Thr
				245				250						255	
Thr	Ala	Ser	Glu	Asp	Ala	Ala	Ala	Leu	Gly	Val	Ala	Ala	Glu	Val	Thr
				260				265					270		
Lys	Glu	Ser	Lys	Val	Val	Lys	Glu	Thr	Glu	Leu	Pro	Val	Glu	Lys	Asn
				275			280					285			
Ala	Val	Ala	Thr	Ala	Cys										
															290

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 292 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide

(B) LOCATION: 1..292

(D) OTHER INFORMATION: / Ceres Seq. ID 1481390

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

Met	Asn	Arg	Gly	Pro	Arg	Ala	Lys	Gly	Phe	Asn	Ser	Gln	Asp	Gly	Ser
1			5					10						15	
Lys	Val	Met	Ala	Val	Ser	Leu	Lys	Glu	Gln	Arg	Val	Thr	Glu	Thr	Glu
			20					25					30		
Lys	Leu	Ser	Glu	Asp	Val	Ser	Leu	Leu	Asp	Pro	Lys	Asp	Tyr	Asn	Lys
			35				40					45			
Ile	Asp	Phe	Pro	Glu	Thr	Tyr	Thr	Glu	Ala	Lys	Phe	Tyr	Val	Ile	Lys
	50					55					60				
Ser	Tyr	Ser	Glu	Asp	Asp	Ile	His	Lys	Ser	Ile	Lys	Tyr	Ser	Val	Trp
65					70					75					80
Ser	Ser	Thr	Pro	Asn	Gly	Asn	Lys	Lys	Leu	Asp	Ala	Ser	Tyr	Asn	Glu
				85					90					95	
Ala	Lys	Gln	Lys	Ser	Asp	Gly	Cys	Pro	Val	Phe	Leu	Leu	Phe	Ser	Val
			100					105					110		
Asn	Thr	Ser	Gly	Gln	Phe	Val	Gly	Leu	Ala	Glu	Met	Val	Gly	Pro	Val
	115						120					125			
Asp	Phe	Asn	Lys	Thr	Val	Glu	Tyr	Trp	Gln	Gln	Asp	Lys	Trp	Ile	Gly
	130					135					140				
Cys	Phe	Pro	Val	Lys	Trp	His	Phe	Val	Lys	Asp	Ile	Pro	Asn	Ser	Ser
145					150					155					160
Leu	Arg	His	Ile	Thr	Leu	Glu	Asn	Asn	Glu	Asn	Lys	Pro	Val	Thr	Asn
			165						170					175	
Ser	Arg	Asp	Thr	Gln	Glu	Val	Lys	Leu	Glu	Gln	Gly	Ile	Lys	Val	Ile
			180					185					190		
Lys	Ile	Phe	Lys	Asp	His	Ala	Ser	Lys	Thr	Cys	Ile	Leu	Asp	Asp	Phe
	195						200					205			
Glu	Phe	Tyr	Glu	Asn	Arg	Gln	Lys	Ile	Ile	Gln	Glu	Arg	Lys	Ser	Lys
	210					215					220				
His	Leu	Gln	Ile	Lys	Lys	Gln	Thr	Leu	Val	Ala	Asn	Ala	Asp	Lys	Gly
225					230						235				240
Val	Met	Ser	Lys	Ile	Asn	Leu	Val	Lys	Pro	Gln	Glu	Ser	Thr	Thr	Ala
			245						250					255	
Ser	Glu	Asp	Ala	Ala	Ala	Leu	Gly	Val	Ala	Ala	Glu	Val	Thr	Lys	Glu
			260					265					270		
Ser	Lys	Val	Val	Lys	Glu	Thr	Glu	Leu	Pro	Val	Glu	Lys	Asn	Ala	Val
	275						280					285			
Ala	Thr	Ala	Cys												
	290														

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 274 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..274

(D) OTHER INFORMATION: / Ceres Seq. ID 1481391

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

Met	Ala	Val	Ser	Leu	Lys	Glu	Gln	Arg	Val	Thr	Glu	Thr	Glu	Lys	Leu
1			5						10					15	
Ser	Glu	Asp	Val	Ser	Leu	Leu	Asp	Pro	Lys	Asp	Tyr	Asn	Lys	Ile	Asp
			20					25					30		
Phe	Pro	Glu	Thr	Tyr	Thr	Glu	Ala	Lys	Phe	Tyr	Val	Ile	Lys	Ser	Tyr
	35						40					45			

Ser Glu Asp Asp Ile His Lys Ser Ile Lys Tyr Ser Val Trp Ser Ser
50 55 60
Thr Pro Asn Gly Asn Lys Lys Leu Asp Ala Ser Tyr Asn Glu Ala Lys
65 70 75 80
Gln Lys Ser Asp Gly Cys Pro Val Phe Leu Phe Ser Val Asn Thr
85 90 95
Ser Gly Gln Phe Val Gly Leu Ala Glu Met Val Gly Pro Val Asp Phe
100 105 110
Asn Lys Thr Val Glu Tyr Trp Gln Gln Asp Lys Trp Ile Gly Cys Phe
115 120 125
Pro Val Lys Trp His Phe Val Lys Asp Ile Pro Asn Ser Ser Leu Arg
130 135 140
His Ile Thr Leu Glu Asn Asn Glu Asn Lys Pro Val Thr Asn Ser Arg
145 150 155 160
Asp Thr Gln Glu Val Lys Leu Glu Gln Gly Ile Lys Val Ile Lys Ile
165 170 175
Phe Lys Asp His Ala Ser Lys Thr Cys Ile Leu Asp Asp Phe Glu Phe
180 185 190
Tyr Glu Asn Arg Gln Lys Ile Ile Gln Glu Arg Lys Ser Lys His Leu
195 200 205
Gln Ile Lys Lys Gln Thr Leu Val Ala Asn Ala Asp Lys Gly Val Met
210 215 220
Ser Lys Ile Asn Leu Val Lys Pro Gln Glu Ser Thr Thr Ala Ser Glu
225 230 235 240
Asp Ala Ala Ala Leu Gly Val Ala Ala Glu Val Thr Lys Glu Ser Lys
245 250 255
Val Val Lys Glu Thr Glu Leu Pro Val Glu Lys Asn Ala Val Ala Thr
260 265 270
Ala Cys

(2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 796 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..796
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481423

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

catcctaatac	gaaaaaaaaagc	aaccaaacac	ataaaaagaga	gattttaatac	aaaagaaaga	60
gaaaaaaaaagaa	agatatggca	ggactcatca	acaagatcgg	agacgcactc	cacaggtcga	120
aggcgaatat	ctcaaagata	tcagaaacgc	caaggatttt	acaatacaca	gcgttgcgaa	180
gtggctcgat	gcacagcttc	cattgcatcc	gcggttgaaa	gctttcttta	ggacaatttc	240
tccgaggcat	tttaaaaaacg	gagattggaa	tacaggtgga	aactgtaaca	acacggttcc	300
tttgtctaga	ggcagcgaaa	tcacagggga	tgatggatcg	atcgatgcaa	cagttgagag	360
tgctgtgaac	gggacaagga	tcaagattct	tgacataact	gcactttctg	agctaagaga	420
cgaagctcat	atctcagggt	ctaaactcaa	accccgaaaa	ccgaagaagg	caagtaacgt	480
gacctcaact	ccaacgatca	acgattgctt	gcattgggtgc	ttaccaggga	tcccagatac	540
ttggaatgaa	cttttcattg	ctcagatttg	aagtattcaa	catcatcaca	cacacaaagc	600
tagctcaatg	gattggctct	gttgattctt	tgttatagaa	aggttttttt	ttcagattct	660
ttcttgggag	aataacaaag	tttcagttct	taaaaatagg	tttttagatgg	tttgtcagta	720
aatgattcat	ctgtaacaat	cacaatctgg	tttttaatta	tacacgagaa	cattgaaatt	780
gaaacaatct	ttttcc					

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 76 amino acids

(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
(A) NAME/KEY: peptide
(B) LOCATION: 1..76
(D) OTHER INFORMATION: / Ceres Seq. ID 1481424

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

```
His Pro Asn Arg Lys Lys Ala Thr Lys His Ile Lys Glu Arg Phe Asn
1          5          10          15
Thr Lys Glu Arg Glu Lys Arg Lys Ile Trp Gln Asp Ser Ser Thr Arg
          20          25          30
Ser Glu Thr His Ser Thr Gly Arg Arg Arg Ile Ser Gln Arg Tyr Gln
          35          40          45
Lys Arg Gln Gly Phe Tyr Asn Thr Gln Arg Cys Glu Val Ala Arg Cys
          50          55          60
Thr Ala Ser Ile Ala Ser Ala Val Glu Ser Phe Leu
          65          70          75
```

(2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 47 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide
(B) LOCATION: 1..47

(D) OTHER INFORMATION: / Ceres Seq. ID 1481425

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

```
Met Ala Gly Leu Ile Asn Lys Ile Gly Asp Ala Leu His Arg Ser Lys
1          5          10          15
Ala Asn Ile Ser Lys Ile Ser Glu Thr Pro Arg Ile Leu Gln Tyr Thr
          20          25          30
Ala Leu Arg Ser Gly Ser Met His Ser Phe His Cys Ile Arg Gly
          35          40          45
```

(2) INFORMATION FOR SEQ ID NO:26:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 492 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..492

(D) OTHER INFORMATION: / Ceres Seq. ID 1481471

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

```
acttcgcctt gaatcgagtc ttcgacgagt ctccggctgc gagtttctct tgctccggca      60
aacagacctg tcattgcttc tctctccggc taactacaca gaagcatggg gtttggacaa      120
gtagtaatag gtccctccagg atcgggaaag accacttatt gcaatggaat gtctcagttc      180
ctctctctaa tgggcaggaa ggttgctatt gttaatctgg atcctgcaaa tgatgcatta      240
ccttatgagt gtgctgtgaa tatagaagaa ttgatcaagt tagaagatgt tatgtcggaa      300
cactcgcttg gtocctaattg aggtcttgta tattgtatgg agtacttgga gaaaaacatt      360
gactggctgg aatctaaact aaagcctctt ctgaaggatc attacattct ctttgatttt      420
cctggccaag tggaattgtt cttcattcat gacagtagca agaattgtct sncgaagctg      480
attaaatcat tg
```

(2) INFORMATION FOR SEQ ID NO:27:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 129 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..129
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481472
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:

Met	Val	Phe	Gly	Gln	Val	Val	Ile	Gly	Pro	Pro	Gly	Ser	Gly	Lys	Thr
1				5				10						15	
Thr	Tyr	Cys	Asn	Gly	Met	Ser	Gln	Phe	Leu	Ser	Leu	Met	Gly	Arg	Lys
			20					25					30		
Val	Ala	Ile	Val	Asn	Leu	Asp	Pro	Ala	Asn	Asp	Ala	Leu	Pro	Tyr	Glu
			35				40					45			
Cys	Ala	Val	Asn	Ile	Glu	Glu	Leu	Ile	Lys	Leu	Glu	Asp	Val	Met	Ser
			50				55				60				
Glu	His	Ser	Leu	Gly	Pro	Asn	Gly	Gly	Leu	Val	Tyr	Cys	Met	Glu	Tyr
				70					75					80	
Leu	Glu	Lys	Asn	Ile	Asp	Trp	Leu	Glu	Ser	Lys	Leu	Lys	Pro	Leu	Leu
			85					90						95	
Lys	Asp	His	Tyr	Ile	Leu	Phe	Asp	Phe	Pro	Gly	Gln	Val	Glu	Leu	Phe
			100					105					110		
Phe	Ile	His	Asp	Ser	Thr	Lys	Asn	Val	Xaa	Xaa	Lys	Leu	Ile	Lys	Ser
			115				120					125			
Leu															

- (2) INFORMATION FOR SEQ ID NO:28:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 108 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: peptide
 (ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..108
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481473
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

Met	Ser	Gln	Phe	Leu	Ser	Leu	Met	Gly	Arg	Lys	Val	Ala	Ile	Val	Asn
1				5				10						15	
Leu	Asp	Pro	Ala	Asn	Asp	Ala	Leu	Pro	Tyr	Glu	Cys	Ala	Val	Asn	Ile
			20					25					30		
Glu	Glu	Leu	Ile	Lys	Leu	Glu	Asp	Val	Met	Ser	Glu	His	Ser	Leu	Gly
			35				40					45			
Pro	Asn	Gly	Gly	Leu	Val	Tyr	Cys	Met	Glu	Tyr	Leu	Glu	Lys	Asn	Ile
			50				55				60				
Asp	Trp	Leu	Glu	Ser	Lys	Leu	Lys	Pro	Leu	Leu	Lys	Asp	His	Tyr	Ile
			70					75						80	
Leu	Phe	Asp	Phe	Pro	Gly	Gln	Val	Glu	Leu	Phe	Phe	Ile	His	Asp	Ser
			85					90					95		
Thr	Lys	Asn	Val	Xaa	Xaa	Lys	Leu	Ile	Lys	Ser	Leu				
			100				105								

- (2) INFORMATION FOR SEQ ID NO:29:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 101 amino acids
 (B) TYPE: amino acid

- (C) STRANDEDNESS:
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
(A) NAME/KEY: peptide
(B) LOCATION: 1..101
(D) OTHER INFORMATION: / Ceres Seq. ID 1481474
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:29:

```
Met Gly Arg Lys Val Ala Ile Val Asn Leu Asp Pro Ala Asn Asp Ala
1           5           10           15
Leu Pro Tyr Glu Cys Ala Val Asn Ile Glu Glu Leu Ile Lys Leu Glu
          20           25           30
Asp Val Met Ser Glu His Ser Leu Gly Pro Asn Gly Gly Leu Val Tyr
          35           40           45
Cys Met Glu Tyr Leu Glu Lys Asn Ile Asp Trp Leu Glu Ser Lys Leu
          50           55           60
Lys Pro Leu Leu Lys Asp His Tyr Ile Leu Phe Asp Phe Pro Gly Gln
65           70           75           80
Val Glu Leu Phe Phe Ile His Asp Ser Thr Lys Asn Val Xaa Xaa Lys
          85           90           95
Leu Ile Lys Ser Leu
          100
```

(2) INFORMATION FOR SEQ ID NO:30:

- (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1189 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: DNA (genomic)
(ix) FEATURE:
(A) NAME/KEY: -
(B) LOCATION: 1..1189
(D) OTHER INFORMATION: / Ceres Seq. ID 1481479

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:

```
cagctcacgg aggaaccagt gttgtctctc aaactccaaa cagaaaaggg agtgtacaca      60
tagctcgctc tcgctctgtg ccccttaacg acaaggaatt aagcctgaag ggaatggatt      120
catttttcag agtaattcct tcgactctc gtgttaagga aggagacgtt ttctcaaagt      180
catcagaggc tggtaatact gaaacagggtg atgctgatgg agaagacata cctgaggatg      240
aagcagtttg taggatttgt ttggtagagc tctgtgaagg aggagaaacc ttaaaaatgg      300
agtgtagttg caaaggcgaa cttgctcttg ccacaaaaga ttgtgctctt aaatggttca      360
ccataaaggg taacaagact tgtgagggtg gtaaacaaaga agttaagaac ttacctgtaa      420
cactcttacg catccaaagc cttcgaaatt ctgggtgttc tcagctagat gtctctggct      480
atagggtgtg gcaggaggta ccggttctag taatcatcag catgctcgct tacttctgct      540
tcctcgagca gtcctgggtt gagaatatgg gtacagggtg catcgctata tcaactgccg      600
tttcttgat tcttggtctt cttgcatcca tgaccgcac aacctaggta atgagaagat      660
ttgtctggat ttacgcatct gtccagtttg cgttggtcgt tctcttcgcc catatatttt      720
actctgtggt gaagttgcaa ccagttctgt cagttcttct gtcaacattt gctggatttg      780
gtgtatgcat atgcggaagt tcagtgatgg ttgagtttgt gagatggaga cgaagatggc      840
gagccagaag gctagagcaa cagctgaacc atgctttgac tctgtcacia ccgccgcaac      900
cactggatcc aacaacctct ctgcatcatt caaataacct atagagagcc aagaagtgga      960
cagatgattt tacatttata cagtgtagtt tgggttaatg ttatgtaatg atttgtataa     1020
aagaaaaaga gaaagtgatc caaggaatgc ttaaaagaty ytccttttgt ttgttttaca     1080
tacacatttg tattgttgta agtttgtaac tttggtttgc tcaatctctg caaatgaaat     1140
gtttgtagca gtattggtt ctctgtataa taaaagatt taaaattgt
```

(2) INFORMATION FOR SEQ ID NO:31:

- (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 313 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:

- (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..313
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481480

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:

Ala	His	Gly	Gly	Thr	Ser	Val	Ala	Pro	Gln	Thr	Pro	Asn	Arg	Lys	Gly
1				5					10					15	
Ser	Val	His	Ile	Ala	Arg	Ser	Arg	Ser	Val	Pro	Leu	Asn	Asp	Lys	Glu
			20					25					30		
Leu	Ser	Leu	Lys	Gly	Met	Asp	Ser	Phe	Phe	Arg	Val	Ile	Pro	Ser	Thr
		35				40						45			
Pro	Arg	Val	Lys	Glu	Gly	Asp	Val	Phe	Ser	Asn	Ala	Ser	Glu	Ala	Gly
	50					55					60				
Asn	Thr	Glu	Thr	Gly	Asp	Ala	Asp	Gly	Glu	Asp	Ile	Pro	Glu	Asp	Glu
65					70				75						80
Ala	Val	Cys	Arg	Ile	Cys	Leu	Val	Glu	Leu	Cys	Glu	Gly	Gly	Glu	Thr
				85					90					95	
Leu	Lys	Met	Glu	Cys	Ser	Cys	Lys	Gly	Glu	Leu	Ala	Leu	Ala	His	Lys
				100				105					110		
Asp	Cys	Ala	Leu	Lys	Trp	Phe	Thr	Ile	Lys	Gly	Asn	Lys	Thr	Cys	Glu
		115					120					125			
Val	Cys	Lys	Gln	Glu	Val	Lys	Asn	Leu	Pro	Val	Thr	Leu	Leu	Arg	Ile
	130					135					140				
Gln	Ser	Leu	Arg	Asn	Ser	Gly	Val	Pro	Gln	Leu	Asp	Val	Ser	Gly	Tyr
145					150					155					160
Arg	Val	Trp	Gln	Glu	Val	Pro	Val	Leu	Val	Ile	Ile	Ser	Met	Leu	Ala
				165					170					175	
Tyr	Phe	Cys	Phe	Leu	Glu	Gln	Leu	Leu	Val	Glu	Asn	Met	Gly	Thr	Gly
			180				185						190		
Ala	Ile	Ala	Ile	Ser	Leu	Pro	Phe	Ser	Cys	Ile	Leu	Gly	Leu	Leu	Ala
		195					200					205			
Ser	Met	Thr	Ala	Ser	Thr	Met	Val	Met	Arg	Arg	Phe	Val	Trp	Ile	Tyr
		210				215					220				
Ala	Ser	Val	Gln	Phe	Ala	Leu	Val	Val	Leu	Phe	Ala	His	Ile	Phe	Tyr
225					230					235					240
Ser	Val	Val	Lys	Leu	Gln	Pro	Val	Leu	Ser	Val	Leu	Leu	Ser	Thr	Phe
				245					250					255	
Ala	Gly	Phe	Gly	Val	Cys	Ile	Cys	Gly	Ser	Ser	Val	Met	Val	Glu	Phe
			260					265					270		
Val	Arg	Trp	Arg	Arg	Arg	Trp	Arg	Ala	Arg	Arg	Leu	Glu	Gln	Gln	Leu
		275					280					285			
Asn	His	Ala	Leu	Thr	Leu	Ser	Gln	Pro	Pro	Gln	Pro	Leu	Asp	Pro	Thr
		290				295					300				
Thr	Ser	Leu	His	His	Ser	Asn	Thr	Ser							
305						310									

(2) INFORMATION FOR SEQ ID NO:32:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 276 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
 (B) LOCATION: 1..276
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481481

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:32:

Met Asp Ser Phe Phe Arg Val Ile Pro Ser Thr Pro Arg Val Lys Glu
1 5 10 15
Gly Asp Val Phe Ser Asn Ala Ser Glu Ala Gly Asn Thr Glu Thr Gly
20 25 30
Asp Ala Asp Gly Glu Asp Ile Pro Glu Asp Glu Ala Val Cys Arg Ile
35 40 45
Cys Leu Val Glu Leu Cys Glu Gly Gly Glu Thr Leu Lys Met Glu Cys
50 55 60
Ser Cys Lys Gly Glu Leu Ala Leu Ala His Lys Asp Cys Ala Leu Lys
65 70 75 80
Trp Phe Thr Ile Lys Gly Asn Lys Thr Cys Glu Val Cys Lys Gln Glu
85 90 95
Val Lys Asn Leu Pro Val Thr Leu Leu Arg Ile Gln Ser Leu Arg Asn
100 105 110
Ser Gly Val Pro Gln Leu Asp Val Ser Gly Tyr Arg Val Trp Gln Glu
115 120 125
Val Pro Val Leu Val Ile Ile Ser Met Leu Ala Tyr Phe Cys Phe Leu
130 135 140
Glu Gln Leu Leu Val Glu Asn Met Gly Thr Gly Ala Ile Ala Ile Ser
145 150 155 160
Leu Pro Phe Ser Cys Ile Leu Gly Leu Leu Ala Ser Met Thr Ala Ser
165 170 175
Thr Met Val Met Arg Arg Phe Val Trp Ile Tyr Ala Ser Val Gln Phe
180 185 190
Ala Leu Val Val Leu Phe Ala His Ile Phe Tyr Ser Val Val Lys Leu
195 200 205
Gln Pro Val Leu Ser Val Leu Leu Ser Thr Phe Ala Gly Phe Gly Val
210 215 220
Cys Ile Cys Gly Ser Ser Val Met Val Glu Phe Val Arg Trp Arg Arg
225 230 235 240
Arg Trp Arg Ala Arg Arg Leu Glu Gln Gln Leu Asn His Ala Leu Thr
245 250 255
Leu Ser Gln Pro Pro Gln Pro Leu Asp Pro Thr Thr Ser Leu His His
260 265 270
Ser Asn Thr Ser
275

(2) INFORMATION FOR SEQ ID NO:33:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 215 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..215
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481482

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:33:

Met Glu Cys Ser Cys Lys Gly Glu Leu Ala Leu Ala His Lys Asp Cys
1 5 10 15
Ala Leu Lys Trp Phe Thr Ile Lys Gly Asn Lys Thr Cys Glu Val Cys
20 25 30
Lys Gln Glu Val Lys Asn Leu Pro Val Thr Leu Leu Arg Ile Gln Ser
35 40 45
Leu Arg Asn Ser Gly Val Pro Gln Leu Asp Val Ser Gly Tyr Arg Val
50 55 60
Trp Gln Glu Val Pro Val Leu Val Ile Ile Ser Met Leu Ala Tyr Phe
65 70 75 80
Cys Phe Leu Glu Gln Leu Leu Val Glu Asn Met Gly Thr Gly Ala Ile

```

      85      90      95
Ala Ile Ser Leu Pro Phe Ser Cys Ile Leu Gly Leu Leu Ala Ser Met
      100      105      110
Thr Ala Ser Thr Met Val Met Arg Arg Phe Val Trp Ile Tyr Ala Ser
      115      120      125
Val Gln Phe Ala Leu Val Val Leu Phe Ala His Ile Phe Tyr Ser Val
      130      135      140
Val Lys Leu Gln Pro Val Leu Ser Val Leu Leu Ser Thr Phe Ala Gly
      145      150      155      160
Phe Gly Val Cys Ile Cys Gly Ser Ser Val Met Val Glu Phe Val Arg
      165      170      175
Trp Arg Arg Arg Trp Arg Ala Arg Arg Leu Glu Gln Gln Leu Asn His
      180      185      190
Ala Leu Thr Leu Ser Gln Pro Pro Gln Pro Leu Asp Pro Thr Thr Ser
      195      200      205
Leu His Ser Asn Thr Ser
      210      215
```

(2) INFORMATION FOR SEQ ID NO:34:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 643 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..643
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481483

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:34:

```

ataattcgag ctgttttttt gctgtaataa tttcacaatt ctcttttttt ttcgctttta      60
aataattttg tctctccatc ttcttctctt ttgctagtct ctcatatcag ctaagaaaag      120
aaattcagaa caaaaaaata acacaaagct ctgtgtttct gtctatctgt tgaatcaa      180
catatggaag acgatcgaaa agagaagaac actccgtggc tatcagtgcc acagtttggt      240
gattgggacc aaaaaggagg aggaacaatg cctgattact ctatggattt cactaagatt      300
agagagatga ggaaacaaaa caagagagac ccttctcgag ccagtttagg caatgaggaa      360
gagctcatta agccaccgca gtcagcaaca tcaactgctg agcttaccac ggtccaaagt      420
gaaaaccgac gagagttctc tcccagccac catcatcaac cacattctcc ttctacgagg      480
agaagtatgt tcagctgctt caactgctgc gttaaagctt gaagatttct tcttgagcaa      540
agtagcagtt ttattattga cttgtgattt gaatgtggaa atgtgttaat gtcatgacac      600
tttaatatat gttccaatcc atttttcttt ttctttggga acc
```

(2) INFORMATION FOR SEQ ID NO:35:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 112 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..112
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481484

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:35:

```

Met Glu Asp Asp Arg Lys Glu Lys Asn Thr Pro Trp Leu Ser Val Pro
  1          5          10          15
Gln Phe Gly Asp Trp Asp Gln Lys Gly Gly Gly Thr Met Pro Asp Tyr
      20      25      30
Ser Met Asp Phe Thr Lys Ile Arg Glu Met Arg Lys Gln Asn Lys Arg
      35      40      45
Asp Pro Ser Arg Ala Ser Leu Gly Asn Glu Glu Glu Leu Ile Lys Pro
```

50	55	60
Pro Glu Ser Ala Thr	Ser Thr Ala Glu Leu Thr	Thr Val Gln Ser Glu
65	70	75
Asn Arg Arg Glu Phe	Ser Pro Ser His His	His Gln Pro His Ser Pro
85	90	95
Ser Thr Arg Arg Ser	Met Phe Ser Cys Phe	Asn Cys Cys Val Lys Ala
100	105	110

(2) INFORMATION FOR SEQ ID NO:36:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 84 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..84
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481485

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:36:

Met	Pro	Asp	Tyr	Ser	Met	Asp	Phe	Thr	Lys	Ile	Arg	Glu	Met	Arg	Lys
1			5					10						15	
Gln	Asn	Lys	Arg	Asp	Pro	Ser	Arg	Ala	Ser	Leu	Gly	Asn	Glu	Glu	Glu
			20					25					30		
Leu	Ile	Lys	Pro	Pro	Glu	Ser	Ala	Thr	Ser	Thr	Ala	Glu	Leu	Thr	Thr
			35				40					45			
Val	Gln	Ser	Glu	Asn	Arg	Arg	Glu	Phe	Ser	Pro	Ser	His	His	His	Gln
			50				55				60				
Pro	His	Ser	Pro	Ser	Thr	Arg	Arg	Ser	Met	Phe	Ser	Cys	Phe	Asn	Cys
65					70				75					80	
Cys	Val	Lys	Ala												

(2) INFORMATION FOR SEQ ID NO:37:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 79 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..79
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481486

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:37:

Met	Asp	Phe	Thr	Lys	Ile	Arg	Glu	Met	Arg	Lys	Gln	Asn	Lys	Arg	Asp
1				5				10						15	
Pro	Ser	Arg	Ala	Ser	Leu	Gly	Asn	Glu	Glu	Leu	Ile	Lys	Pro	Pro	
			20					25					30		
Glu	Ser	Ala	Thr	Ser	Thr	Ala	Glu	Leu	Thr	Thr	Val	Gln	Ser	Glu	Asn
			35				40					45			
Arg	Arg	Glu	Phe	Ser	Pro	Ser	His	His	His	Gln	Pro	His	Ser	Pro	Ser
			50				55				60				
Thr	Arg	Arg	Ser	Met	Phe	Ser	Cys	Phe	Asn	Cys	Cys	Val	Lys	Ala	
65				70					75						

(2) INFORMATION FOR SEQ ID NO:38:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 760 base pairs
- (B) TYPE: nucleic acid

- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- (ix) FEATURE:
 - (A) NAME/KEY: -
 - (B) LOCATION: 1..760
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481487
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:38:

```
gcmccmbtyy cattayytag aacatcctaw hraaaaaaca aaagtgatca gttttgtttt    60
ctcgggggaaa ttttctgaaa gtgaagaaag ggaaagcaag ttttttttga agtgggggaga    120
gagatgggag aaatggggaa ggcgatggga ttgctgatta gcgggacgct tgtgtattac    180
cattgtgcat atcgtaacgc gactcttctc tctctcttct ccgatgtttt cattgttctc    240
ttatgctctc tcgccattct cggctctcctt tttcgccaac tcaatgtctc ggtaccagtg    300
gatccactag agtggcaaat atcacaggac acagcaagta acatcgttgc acgcttagct    360
aataccggtt gagcagcaga ggggtgttctg agggttgcag caactggaca tgacaagaga    420
ctttttgtca aggtcgtaat ttgcctttac ttcttatcag cgcttgggag actcatatca    480
ggggttaaccg ttgcttatgc aggactatgc ttgttctgtc tctccatgct ctgtcagact    540
tctcaatctc ttggaaactg tgtactaaag cgaggaaatg gccagatttt agaacaagaa    600
gcacattctg atacataata tgtctagctt ttgtttatac ttttcgtctt ttctcatgct    660
tacatgctca tagcttcagt cttcagagta gtttcccctt atgtacattg gatttgttgc    720
atactacctt gtgaaaaatg taatgatatt gtttaacctc
```

(2) INFORMATION FOR SEQ ID NO:39:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 164 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..164
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481488
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:39:

```
Met Gly Glu Met Gly Lys Ala Met Gly Leu Leu Ile Ser Gly Thr Leu
1          5          10          15
Val Tyr Tyr His Cys Ala Tyr Arg Asn Ala Thr Leu Leu Ser Leu Phe
20          25          30
Ser Asp Val Phe Ile Val Leu Leu Cys Ser Leu Ala Ile Leu Gly Leu
35          40          45
Leu Phe Arg Gln Leu Asn Val Ser Val Pro Val Asp Pro Leu Glu Trp
50          55          60
Gln Ile Ser Gln Asp Thr Ala Ser Asn Ile Val Ala Arg Leu Ala Asn
65          70          75          80
Thr Val Gly Ala Ala Glu Gly Val Leu Arg Val Ala Ala Thr Gly His
85          90          95
Asp Lys Arg Leu Phe Val Lys Val Val Ile Cys Leu Tyr Phe Leu Ser
100         105         110
Ala Leu Gly Arg Leu Ile Ser Xaa Val Thr Val Ala Tyr Ala Gly Leu
115         120         125
Cys Leu Phe Cys Leu Ser Met Leu Cys Gln Thr Ser Gln Ser Leu Gly
130         135         140
Asn Cys Val Leu Lys Arg Gly Asn Gly Gln Ile Leu Glu Gln Glu Ala
145         150         155         160
His Ser Asp Thr
```

(2) INFORMATION FOR SEQ ID NO:40:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 161 amino acids
 - (B) TYPE: amino acid

- (C) STRANDEDNESS:
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
(A) NAME/KEY: peptide
(B) LOCATION: 1..161
(D) OTHER INFORMATION: / Ceres Seq. ID 1481489
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:40:

Met	Gly	Lys	Ala	Met	Gly	Leu	Leu	Ile	Ser	Gly	Thr	Leu	Val	Tyr	Tyr
1				5					10					15	
His	Cys	Ala	Tyr	Arg	Asn	Ala	Thr	Leu	Leu	Ser	Leu	Phe	Ser	Asp	Val
			20					25					30		
Phe	Ile	Val	Leu	Leu	Cys	Ser	Leu	Ala	Ile	Leu	Gly	Leu	Leu	Phe	Arg
		35				40						45			
Gln	Leu	Asn	Val	Ser	Val	Pro	Val	Asp	Pro	Leu	Glu	Trp	Gln	Ile	Ser
	50					55					60				
Gln	Asp	Thr	Ala	Ser	Asn	Ile	Val	Ala	Arg	Leu	Ala	Asn	Thr	Val	Gly
65					70				75					80	
Ala	Ala	Glu	Gly	Val	Leu	Arg	Val	Ala	Ala	Thr	Gly	His	Asp	Lys	Arg
				85				90						95	
Leu	Phe	Val	Lys	Val	Val	Ile	Cys	Leu	Tyr	Phe	Leu	Ser	Ala	Leu	Gly
			100					105					110		
Arg	Leu	Ile	Ser	Xaa	Val	Thr	Val	Ala	Tyr	Ala	Gly	Leu	Cys	Leu	Phe
		115					120						125		
Cys	Leu	Ser	Met	Leu	Cys	Gln	Thr	Ser	Gln	Ser	Leu	Gly	Asn	Cys	Val
	130					135					140				
Leu	Lys	Arg	Gly	Asn	Gly	Gln	Ile	Leu	Glu	Gln	Glu	Ala	His	Ser	Asp
145					150					155					160
Thr															

- (2) INFORMATION FOR SEQ ID NO:41:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 157 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
(A) NAME/KEY: peptide
(B) LOCATION: 1..157
(D) OTHER INFORMATION: / Ceres Seq. ID 1481490
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:41:

Met	Gly	Leu	Leu	Ile	Ser	Gly	Thr	Leu	Val	Tyr	Tyr	His	Cys	Ala	Tyr
1				5					10					15	
Arg	Asn	Ala	Thr	Leu	Leu	Ser	Leu	Phe	Ser	Asp	Val	Phe	Ile	Val	Leu
			20					25					30		
Leu	Cys	Ser	Leu	Ala	Ile	Leu	Gly	Leu	Leu	Phe	Arg	Gln	Leu	Asn	Val
		35				40						45			
Ser	Val	Pro	Val	Asp	Pro	Leu	Glu	Trp	Gln	Ile	Ser	Gln	Asp	Thr	Ala
	50					55					60				
Ser	Asn	Ile	Val	Ala	Arg	Leu	Ala	Asn	Thr	Val	Gly	Ala	Ala	Glu	Gly
65					70				75					80	
Val	Leu	Arg	Val	Ala	Ala	Thr	Gly	His	Asp	Lys	Arg	Leu	Phe	Val	Lys
				85				90						95	
Val	Val	Ile	Cys	Leu	Tyr	Phe	Leu	Ser	Ala	Leu	Gly	Arg	Leu	Ile	Ser
			100					105					110		
Xaa	Val	Thr	Val	Ala	Tyr	Ala	Gly	Leu	Cys	Leu	Phe	Cys	Leu	Ser	Met
		115					120					125			
Leu	Cys	Gln	Thr	Ser	Gln	Ser	Leu	Gly	Asn	Cys	Val	Leu	Lys	Arg	Gly

130 135 140
Asn Gly Gln Ile Leu Glu Gln Glu Ala His Ser Asp Thr
145 150 155

(2) INFORMATION FOR SEQ ID NO:42:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 661 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..661
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481491

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:42:

```
mcacaaaaya actaaaaaac aatcagatct gagatcgaac aaaacaacat gaacacgtta      60
atcccatcgg agaaaaagatg gatcatcacc ggcgttttac tagccgggtt agttggcggt      120
gctttgcttt tcacaagctt catacgagcc gctgacgaaa cgctcttcct ctgttccaca      180
gcaagcgcca aaagcagagc ggtggctgcg gcagctgatt acgaagcgac tccgattcag      240
cttcaagcga tcgtccacta cgcgacatct aacgttgtcc cacaacagaa tcttgctgag      300
atctcgatct ctttcaacat cttgaaaaag ctagctccgg ctaactttct cgtgttcggt      360
ctcggctcgt actcgctcat gtgggcttct ttaaattccac gtggcaaaac cttgttcttg      420
gaagaagatc ttgaatggtt tcagaaagtg accaaagact ctcctttctt acgtgcgcac      480
cacgtgcgtt acaggacgca gcttcaacaa gccgattcgc ttctacgttc gtacaaaacg      540
gagcctaact gttttccggc gaaatcttat ctccggggaa acgagaagtg taagctagct      600
ctcacgggac tgcccgatga gttctacgat acagagtggg atctgctgat ggtcgatgct      660
```

(2) INFORMATION FOR SEQ ID NO:43:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 204 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..204
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481492

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:43:

```
Met Asn Thr Leu Ile Pro Ser Glu Lys Arg Trp Ile Ile Thr Gly Val
1      5      10      15
Leu Leu Ala Gly Leu Val Gly Gly Ala Leu Leu Phe Thr Ser Phe Ile
20     25     30
Arg Ala Ala Asp Glu Thr Leu Phe Leu Cys Ser Thr Ala Ser Ala Lys
35     40     45
Ser Arg Ala Val Ala Ala Ala Ala Asp Tyr Glu Ala Thr Pro Ile Gln
50     55     60
Leu Gln Ala Ile Val His Tyr Ala Thr Ser Asn Val Val Pro Gln Gln
65     70     75     80
Asn Leu Ala Glu Ile Ser Ile Ser Phe Asn Ile Leu Lys Lys Leu Ala
85     90     95
Pro Ala Asn Phe Leu Val Phe Gly Leu Gly Arg Asp Ser Leu Met Trp
100    105    110
Ala Ser Leu Asn Pro Arg Gly Lys Thr Leu Phe Leu Glu Glu Asp Leu
115    120    125
Glu Trp Phe Gln Lys Val Thr Lys Asp Ser Pro Phe Leu Arg Ala His
130    135    140
His Val Arg Tyr Arg Thr Gln Leu Gln Gln Ala Asp Ser Leu Leu Arg
145    150    155    160
```

Ser Tyr Lys Thr Glu Pro Asn Cys Phe Pro Ala Lys Ser Tyr Leu Arg
165 170 175
Gly Asn Glu Lys Cys Lys Leu Ala Leu Thr Gly Leu Pro Asp Glu Phe
180 185 190
Tyr Asp Thr Glu Trp Asp Leu Leu Met Val Asp Ala
195 200

(2) INFORMATION FOR SEQ ID NO:44:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1163 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1163
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481504

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:44:

aatgctcgta agttcaagca aaatcacaag agcgagagag atgggtgacga aaacagagga	60
gaagcaattg aaccagctag agattcaagt cgataatggc ggaggtggaa catgggagta	120
tctttgtctc gttcgtaatc tcaaacttcg tcggtcggag aaagtattaa aacacggttc	180
ctcgattttg aatgatccga ggaaacgacg tgctctcggg ccatatgaat ggacactaaa	240
tgagcagggtg gcaattgcag ctatggactg tcaatgtctc ggtgtcgcac agagttgcat	300
taaggctttg cagaagaaat ttcctgggag caaaagggtt gggaggcttg aggcattgct	360
tcttgaagca aagggattat ggggagaggc tgaggaagca tatgcgagtc ttttgggaaga	420
taatccactc gaccaagcga tacacaaacg aagagtggct atatccaagg cactaggaaa	480
accttccata gccattgagc ttcttaacaa atatcttgaa ctattcatgg ctgatcatga	540
tgcatggaga gaacttgcag agctttatct ttccttgcaa atgtataagc aagcagcttt	600
ctgctatgaa gagctcatac tatctcagcc tactgttcca ttgtaccacc tcgcatatgc	660
tgagggttctc tatacaatcg gtggagtaga aaacattatc tcagcaagaa aatactatgc	720
agcgaccgta gatttaacag gcggcaaaaa cactagagct cttctcggaa tctgcttggtg	780
tgcatcggcc attgcacagc tctcaaaagg caggacaaa gaggacaaag acgctacggc	840
agccccagag cttcattccc tggctgcagc tgcagttagg aaagaatata agcaaaaagc	900
cccggacaaa cttcagctca tctcttcgag gttaagaatc ttgaagactt gatcgcaagt	960
aaacgatggt ctggccacac agacgcaaac gacttagcag tagtagatag tcggaaaata	1020
tcgaactcta aattcaata actttcttta aagtttaaac caaagagaat tttgattact	1080
gtagatatac aaaccaaata aactgtatca ctccttagcc tttacggttt ccattgctgc	1140
gacgtgcagc ttcttttgta tcg	

(2) INFORMATION FOR SEQ ID NO:45:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 316 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..316
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481505

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:45:

Met Leu Val Ser Ser Ser Lys Ile Thr Arg Ala Arg Glu Met Val Thr	
1 5 10 15	
Lys Thr Glu Glu Lys Gln Leu Asn Gln Leu Glu Ile Gln Val Asp Asn	
20 25 30	
Gly Gly Gly Gly Thr Trp Glu Tyr Leu Cys Leu Val Arg Asn Leu Lys	
35 40 45	
Leu Arg Arg Ser Glu Lys Val Leu Lys His Gly Ser Ser Ile Leu Asn	
50 55 60	
Asp Pro Arg Lys Arg Ser Ala Leu Gly Pro Tyr Glu Trp Thr Leu Asn	

65					70					75					80
Glu	Gln	Val	Ala	Ile	Ala	Ala	Met	Asp	Cys	Gln	Cys	Leu	Gly	Val	Ala
					85				90					95	
Gln	Ser	Cys	Ile	Lys	Ala	Leu	Gln	Lys	Lys	Phe	Pro	Gly	Ser	Lys	Arg
			100					105					110		
Val	Gly	Arg	Leu	Glu	Ala	Leu	Leu	Glu	Ala	Lys	Gly	Leu	Trp	Gly	
			115				120					125			
Glu	Ala	Glu	Glu	Ala	Tyr	Ala	Ser	Leu	Leu	Glu	Asp	Asn	Pro	Leu	Asp
			130				135				140				
Gln	Ala	Ile	His	Lys	Arg	Arg	Val	Ala	Ile	Ser	Lys	Ala	Leu	Gly	Lys
145					150					155				160	
Pro	Ser	Ile	Ala	Ile	Glu	Leu	Leu	Asn	Lys	Tyr	Leu	Glu	Leu	Phe	Met
				165					170					175	
Ala	Asp	His	Asp	Ala	Trp	Arg	Glu	Leu	Ala	Glu	Leu	Tyr	Leu	Ser	Leu
			180					185					190		
Gln	Met	Tyr	Lys	Gln	Ala	Ala	Phe	Cys	Tyr	Glu	Glu	Leu	Ile	Leu	Ser
			195				200					205			
Gln	Pro	Thr	Val	Pro	Leu	Tyr	His	Leu	Ala	Tyr	Ala	Glu	Val	Leu	Tyr
			210				215					220			
Thr	Ile	Gly	Gly	Val	Glu	Asn	Ile	Ile	Ser	Ala	Arg	Lys	Tyr	Tyr	Ala
225					230					235				240	
Ala	Thr	Val	Asp	Leu	Thr	Gly	Gly	Lys	Asn	Thr	Arg	Ala	Leu	Leu	Gly
				245					250					255	
Ile	Cys	Leu	Cys	Ala	Ser	Ala	Ile	Ala	Gln	Leu	Ser	Lys	Gly	Arg	Asn
			260					265					270		
Lys	Glu	Asp	Lys	Asp	Ala	Thr	Ala	Ala	Pro	Glu	Leu	His	Ser	Leu	Ala
			275				280					285			
Ala	Ala	Ala	Val	Glu	Lys	Glu	Tyr	Lys	Gln	Lys	Ala	Pro	Asp	Lys	Leu
			290				295				300				
Gln	Leu	Ile	Ser	Ser	Ala	Leu	Arg	Ile	Leu	Lys	Thr				
305					310					315					

(2) INFORMATION FOR SEQ ID NO:46:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 316 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..316
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481506

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:46:

Met	Leu	Val	Ser	Ser	Ser	Lys	Ile	Thr	Arg	Ala	Arg	Glu	Met	Val	Thr
1				5				10						15	
Lys	Thr	Glu	Glu	Lys	Gln	Leu	Asn	Gln	Leu	Glu	Ile	Gln	Val	Asp	Asn
			20					25					30		
Gly	Gly	Gly	Gly	Thr	Trp	Glu	Tyr	Leu	Cys	Leu	Val	Arg	Asn	Leu	Lys
			35				40					45			
Leu	Arg	Arg	Ser	Glu	Lys	Val	Leu	Lys	His	Gly	Ser	Ser	Ile	Leu	Asn
			50				55				60				
Asp	Pro	Arg	Lys	Arg	Ser	Ala	Leu	Gly	Pro	Tyr	Glu	Trp	Thr	Leu	Asn
65					70				75					80	
Glu	Gln	Val	Ala	Ile	Ala	Ala	Met	Asp	Cys	Gln	Cys	Leu	Gly	Val	Ala
			85						90					95	
Gln	Ser	Cys	Ile	Lys	Ala	Leu	Gln	Lys	Lys	Phe	Pro	Gly	Ser	Lys	Arg
			100					105					110		
Val	Gly	Arg	Leu	Glu	Ala	Leu	Leu	Glu	Ala	Lys	Gly	Leu	Trp	Gly	
			115				120					125			

Glu Ala Glu Glu Ala Tyr Ala Ser Leu Leu Glu Asp Asn Pro Leu Asp
130 135 140
Gln Ala Ile His Lys Arg Arg Val Ala Ile Ser Lys Ala Leu Gly Lys
145 150 155 160
Pro Ser Ile Ala Ile Glu Leu Leu Asn Lys Tyr Leu Glu Leu Phe Met
165 170 175
Ala Asp His Asp Ala Trp Arg Glu Leu Ala Glu Leu Tyr Leu Ser Leu
180 185 190
Gln Met Tyr Lys Gln Ala Ala Phe Cys Tyr Glu Glu Leu Ile Leu Ser
195 200 205
Gln Pro Thr Val Pro Leu Tyr His Leu Ala Tyr Ala Glu Val Leu Tyr
210 215 220
Thr Ile Gly Gly Val Glu Asn Ile Ile Ser Ala Arg Lys Tyr Tyr Ala
225 230 235 240
Ala Thr Val Asp Leu Thr Gly Gly Lys Asn Thr Arg Ala Leu Leu Gly
245 250 255
Ile Cys Leu Cys Ala Ser Ala Ile Ala Gln Leu Ser Lys Gly Arg Asn
260 265 270
Lys Glu Asp Lys Asp Ala Thr Ala Ala Pro Glu Leu His Ser Leu Ala
275 280 285
Ala Ala Ala Val Glu Lys Glu Tyr Lys Gln Lys Ala Pro Asp Lys Leu
290 295 300
Gln Leu Ile Ser Ser Ala Leu Arg Ile Leu Lys Thr
305 310 315

(2) INFORMATION FOR SEQ ID NO:47:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 303 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..303

(D) OTHER INFORMATION: / Ceres Seq. ID 1481507

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:47:

Met Val Thr Lys Thr Glu Glu Lys Gln Leu Asn Gln Leu Glu Ile Gln
1 5 10 15
Val Asp Asn Gly Gly Gly Gly Thr Trp Glu Tyr Leu Cys Leu Val Arg
20 25 30
Asn Leu Lys Leu Arg Arg Ser Glu Lys Val Leu Lys His Gly Ser Ser
35 40 45
Ile Leu Asn Asp Pro Arg Lys Arg Ser Ala Leu Gly Pro Tyr Glu Trp
50 55 60
Thr Leu Asn Glu Gln Val Ala Ile Ala Ala Met Asp Cys Gln Cys Leu
65 70 75 80
Gly Val Ala Gln Ser Cys Ile Lys Ala Leu Gln Lys Lys Phe Pro Gly
85 90 95
Ser Lys Arg Val Gly Arg Leu Glu Ala Leu Leu Glu Ala Lys Gly
100 105 110
Leu Trp Gly Glu Ala Glu Glu Ala Tyr Ala Ser Leu Leu Glu Asp Asn
115 120 125
Pro Leu Asp Gln Ala Ile His Lys Arg Arg Val Ala Ile Ser Lys Ala
130 135 140
Leu Gly Lys Pro Ser Ile Ala Ile Glu Leu Leu Asn Lys Tyr Leu Glu
145 150 155 160
Leu Phe Met Ala Asp His Asp Ala Trp Arg Glu Leu Ala Glu Leu Tyr
165 170 175
Leu Ser Leu Gln Met Tyr Lys Gln Ala Ala Phe Cys Tyr Glu Glu Leu

	180		185		190										
Ile	Leu	Ser	Gln	Pro	Thr	Val	Pro	Leu	Tyr	His	Leu	Ala	Tyr	Ala	Glu
	195						200					205			
Val	Leu	Tyr	Thr	Ile	Gly	Gly	Val	Glu	Asn	Ile	Ile	Ser	Ala	Arg	Lys
	210					215					220				
Tyr	Tyr	Ala	Ala	Thr	Val	Asp	Leu	Thr	Gly	Gly	Lys	Asn	Thr	Arg	Ala
225					230					235					240
Leu	Leu	Gly	Ile	Cys	Leu	Cys	Ala	Ser	Ala	Ile	Ala	Gln	Leu	Ser	Lys
				245					250					255	
Gly	Arg	Asn	Lys	Glu	Asp	Lys	Asp	Ala	Thr	Ala	Ala	Pro	Glu	Leu	His
			260					265					270		
Ser	Leu	Ala	Ala	Ala	Ala	Val	Glu	Lys	Glu	Tyr	Lys	Gln	Lys	Ala	Pro
	275						280					285			
Asp	Lys	Leu	Gln	Leu	Ile	Ser	Ser	Ala	Leu	Arg	Ile	Leu	Lys	Thr	
	290					295					300				

(2) INFORMATION FOR SEQ ID NO:48:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1259 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1259
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481516

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:48:

atctcacttt	ccgattttat	aaaattgatt	cttctcttct	tcttaaacc	atgaagagtt	60
catgatttct	taagctcgca	gcataatcga	tggcgaattt	gagtttgagc	ttgtatctaa	120
tcctccggat	ttacgctctt	ttgttgctgt	tcaatgtctc	cttcgctaaa	acacttaaac	180
gagacatgaa	agctttgaat	gagataaaga	aattggtggg	atggagattg	gtatactctt	240
gggttgagga	tgatccttgt	ggcgatggag	ttttgcctcc	gtggtctgga	gttacttgct	300
ctaaagttgg	cgattatcgt	gtcgtcgtca	agctagaagt	gtattcaatg	tcgatagttg	360
ggaatttccc	aaaggctata	acgaagctct	tagatctcac	tgttttggat	atgcataata	420
acaaattaac	aggtcctatt	cctccagaaa	ttgggcggct	taagcggctt	atcacactga	480
atctgaggtg	gaacaaactt	caacaggcac	tgctcctga	aattggtgga	ttgaagagtc	540
taacttatct	gtacctgagt	tttaacaatt	tcaaaggaga	aatccccaaa	gaacttgcaa	600
atctccatga	gctccagtac	ttacatattc	aggagaatca	ttttactggg	cgaattccag	660
cagagctggg	aacattacaa	aaacttcgcc	acttggatgc	tggcaacaat	aacttagtgg	720
ggagtataag	cgatcttttt	cgcattgaag	gatgctttcc	agctcttaga	aacctgtttt	780
taaacaataa	ttacttgact	ggaggactcc	caaacaagct	tgcaaactca	acaaacctgg	840
agatcttgta	cttatctttc	aacaaaatga	ctggagcaat	acccgctgca	cttgccagta	900
taccaagact	aactaacttg	cacttggaac	acaatctatt	caatggaagt	atacctgaag	960
ccttctacaa	gcattccaaac	ctaaaagata	tgtacataga	agggaatgct	ttcaaatacag	1020
acgtgaaggc	gattgggtgca	cataaggtcc	tcgaactttc	tgacacagac	ttccttggtt	1080
agttatgtat	agcacaactt	tgtttcattt	acagatagga	atttggcagt	gttatctggt	1140
tatttaagat	tcattttctc	tgttaaagcg	agattgtagt	tgatgtgttt	tctgaatgta	1200
aaagattcct	tatccatgta	tgaaaattga	atataaaggg	aatctgggtt	gttctttcc	

(2) INFORMATION FOR SEQ ID NO:49:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 330 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..330
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481517

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:49:

Met	Ala	Asn	Leu	Ser	Leu	Ser	Leu	Tyr	Leu	Ile	Leu	Arg	Ile	Tyr	Ala
1				5				10						15	
Leu	Leu	Leu	Leu	Phe	Asn	Val	Ser	Phe	Ala	Lys	Thr	Leu	Lys	Arg	Asp
			20					25					30		
Met	Lys	Ala	Leu	Asn	Glu	Ile	Lys	Lys	Leu	Val	Gly	Trp	Arg	Leu	Val
		35					40					45			
Tyr	Ser	Trp	Val	Gly	Asp	Asp	Pro	Cys	Gly	Asp	Gly	Val	Leu	Pro	Pro
	50					55					60				
Trp	Ser	Gly	Val	Thr	Cys	Ser	Lys	Val	Gly	Asp	Tyr	Arg	Val	Val	Val
65					70					75					80
Lys	Leu	Glu	Val	Tyr	Ser	Met	Ser	Ile	Val	Gly	Asn	Phe	Pro	Lys	Ala
				85					90					95	
Ile	Thr	Lys	Leu	Leu	Asp	Leu	Thr	Val	Leu	Asp	Met	His	Asn	Asn	Lys
			100					105					110		
Leu	Thr	Gly	Pro	Ile	Pro	Pro	Glu	Ile	Gly	Arg	Leu	Lys	Arg	Leu	Ile
		115					120					125			
Thr	Leu	Asn	Leu	Arg	Trp	Asn	Lys	Leu	Gln	Gln	Ala	Leu	Pro	Pro	Glu
	130					135					140				
Ile	Gly	Gly	Leu	Lys	Ser	Leu	Thr	Tyr	Leu	Tyr	Leu	Ser	Phe	Asn	Asn
145					150					155					160
Phe	Lys	Gly	Glu	Ile	Pro	Lys	Glu	Leu	Ala	Asn	Leu	His	Glu	Leu	Gln
			165						170					175	
Tyr	Leu	His	Ile	Gln	Glu	Asn	His	Phe	Thr	Gly	Arg	Ile	Pro	Ala	Glu
			180					185					190		
Leu	Gly	Thr	Leu	Gln	Lys	Leu	Arg	His	Leu	Asp	Ala	Gly	Asn	Asn	Asn
		195					200				205				
Leu	Val	Gly	Ser	Ile	Ser	Asp	Leu	Phe	Arg	Ile	Glu	Gly	Cys	Phe	Pro
	210					215					220				
Ala	Leu	Arg	Asn	Leu	Phe	Leu	Asn	Asn	Asn	Tyr	Leu	Thr	Gly	Gly	Leu
225					230					235					240
Pro	Asn	Lys	Leu	Ala	Asn	Leu	Thr	Asn	Leu	Glu	Ile	Leu	Tyr	Leu	Ser
			245						250					255	
Phe	Asn	Lys	Met	Thr	Gly	Ala	Ile	Pro	Ala	Ala	Leu	Ala	Ser	Ile	Pro
			260					265					270		
Arg	Leu	Thr	Asn	Leu	His	Leu	Asp	His	Asn	Leu	Phe	Asn	Gly	Ser	Ile
		275					280						285		
Pro	Glu	Ala	Phe	Tyr	Lys	His	Pro	Asn	Leu	Lys	Asp	Met	Tyr	Ile	Glu
	290					295					300				
Gly	Asn	Ala	Phe	Lys	Ser	Asp	Val	Lys	Ala	Ile	Gly	Ala	His	Lys	Val
305					310					315					320
Leu	Glu	Leu	Ser	Asp	Thr	Asp	Phe	Leu	Val						
			325					330							

(2) INFORMATION FOR SEQ ID NO:50:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 298 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..298

(D) OTHER INFORMATION: / Ceres Seq. ID 1481518

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:50:

Met	Lys	Ala	Leu	Asn	Glu	Ile	Lys	Lys	Leu	Val	Gly	Trp	Arg	Leu	Val
1				5					10					15	
Tyr	Ser	Trp	Val	Gly	Asp	Asp	Pro	Cys	Gly	Asp	Gly	Val	Leu	Pro	Pro
		20						25					30		

Trp Ser Gly Val Thr Cys Ser Lys Val Gly Asp Tyr Arg Val Val Val
35 40 45
Lys Leu Glu Val Tyr Ser Met Ser Ile Val Gly Asn Phe Pro Lys Ala
50 55 60
Ile Thr Lys Leu Leu Asp Leu Thr Val Leu Asp Met His Asn Asn Lys
65 70 75 80
Leu Thr Gly Pro Ile Pro Pro Glu Ile Gly Arg Leu Lys Arg Leu Ile
85 90 95
Thr Leu Asn Leu Arg Trp Asn Lys Leu Gln Gln Ala Leu Pro Pro Glu
100 105 110
Ile Gly Gly Leu Lys Ser Leu Thr Tyr Leu Tyr Leu Ser Phe Asn Asn
115 120 125
Phe Lys Gly Glu Ile Pro Lys Glu Leu Ala Asn Leu His Glu Leu Gln
130 135 140
Tyr Leu His Ile Gln Glu Asn His Phe Thr Gly Arg Ile Pro Ala Glu
145 150 155 160
Leu Gly Thr Leu Gln Lys Leu Arg His Leu Asp Ala Gly Asn Asn Asn
165 170 175
Leu Val Gly Ser Ile Ser Asp Leu Phe Arg Ile Glu Gly Cys Phe Pro
180 185 190
Ala Leu Arg Asn Leu Phe Leu Asn Asn Asn Tyr Leu Thr Gly Gly Leu
195 200 205
Pro Asn Lys Leu Ala Asn Leu Thr Asn Leu Glu Ile Leu Tyr Leu Ser
210 215 220
Phe Asn Lys Met Thr Gly Ala Ile Pro Ala Ala Leu Ala Ser Ile Pro
225 230 235 240
Arg Leu Thr Asn Leu His Leu Asp His Asn Leu Phe Asn Gly Ser Ile
245 250 255
Pro Glu Ala Phe Tyr Lys His Pro Asn Leu Lys Asp Met Tyr Ile Glu
260 265 270
Gly Asn Ala Phe Lys Ser Asp Val Lys Ala Ile Gly Ala His Lys Val
275 280 285
Leu Glu Leu Ser Asp Thr Asp Phe Leu Val
290 295

(2) INFORMATION FOR SEQ ID NO:51:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 244 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..244

(D) OTHER INFORMATION: / Ceres Seq. ID 1481519

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:51:

Met Ser Ile Val Gly Asn Phe Pro Lys Ala Ile Thr Lys Leu Leu Asp
1 5 10 15
Leu Thr Val Leu Asp Met His Asn Asn Lys Leu Thr Gly Pro Ile Pro
20 25 30
Pro Glu Ile Gly Arg Leu Lys Arg Leu Ile Thr Leu Asn Leu Arg Trp
35 40 45
Asn Lys Leu Gln Gln Ala Leu Pro Pro Glu Ile Gly Gly Leu Lys Ser
50 55 60
Leu Thr Tyr Leu Tyr Leu Ser Phe Asn Asn Phe Lys Gly Glu Ile Pro
65 70 75 80
Lys Glu Leu Ala Asn Leu His Glu Leu Gln Tyr Leu His Ile Gln Glu
85 90 95
Asn His Phe Thr Gly Arg Ile Pro Ala Glu Leu Gly Thr Leu Gln Lys

	100		105		110
Leu Arg His	Leu Asp Ala Gly Asn Asn Asn Leu Val Gly Ser Ile Ser				
115			120		125
Asp Leu Phe Arg Ile Glu Gly Cys Phe Pro Ala Leu Arg Asn Leu Phe					
130			135		140
Leu Asn Asn Asn Tyr Leu Thr Gly Gly Leu Pro Asn Lys Leu Ala Asn					
145			150		155
Leu Thr Asn Leu Glu Ile Leu Tyr Leu Ser Phe Asn Lys Met Thr Gly					
	165		170		175
Ala Ile Pro Ala Ala Leu Ala Ser Ile Pro Arg Leu Thr Asn Leu His					
	180		185		190
Leu Asp His Asn Leu Phe Asn Gly Ser Ile Pro Glu Ala Phe Tyr Lys					
195			200		205
His Pro Asn Leu Lys Asp Met Tyr Ile Glu Gly Asn Ala Phe Lys Ser					
210			215		220
Asp Val Lys Ala Ile Gly Ala His Lys Val Leu Glu Leu Ser Asp Thr					
225			230		235
Asp Phe Leu Val					240

(2) INFORMATION FOR SEQ ID NO:52:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 860 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..860
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481520

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:52:

cattaagctg actaagttcg agaacgagga agctgtctgc aacccccaaa gaactcgtgc	60
taatgatatg aagaatttag ccactgctgc tgtaaaagca agcagatttt atagggagtt	120
gaattcccaa actgtcaaac acttgacac actccatgag taccttggca tgatgatggc	180
tgtccaaggc gcatttgcag atagatctag tgctttactg acagttcaga cgcttctatc	240
agagcttctt tctctgcaaa ctagagttag gaagctagag gctgcacatc cgaaggattt	300
tggtggtgac aaatcaagga tccgaaaaat agaagagtta aaagaaacaa tcaaggtcac	360
tgaggatgca aaaaatggtt ccatcaaagg gtatgagcga atcaaggaaa acaaccgato	420
tgaggttgag aggttggaca gagaaaggcg tgcagacttc atgaacatga tgaagggttt	480
tggtgttaac caggttggat acgcagagaa aatgggaaac gtctgggcaa aggttgcaga	540
agagaccagc caatacgata gagagaagca gacagctaa caaacacaga aaaaagaga	600
gtgaacgatg ttcatttttg cataaccata ccaaattccat gtatggcaca gaatcacatt	660
gcgtaataat ggtttgtcaa aaagtgtagt ttcctttttc atatgttgta tctatcttga	720
tagagatttg taaacgttct tggttgtttt cttagttgac tgtaaattag ttttctagaa	780
gcattctctg ctacagctgc attgactcat acccattgtt ttctggtata tgccgcaaaa	840
gatatatctg atagtttggc	

(2) INFORMATION FOR SEQ ID NO:53:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 192 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..192
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481521

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:53:

Ile Lys Leu Thr Lys Phe Glu Asn Glu Glu Ala Val Cys Asn Pro Gln

1				5					10					15		
Arg	Thr	Arg	Ala	Asn	Asp	Met	Lys	Asn	Leu	Ala	Thr	Ala	Ala	Val	Lys	
			20					25					30			
Ala	Ser	Arg	Phe	Tyr	Arg	Glu	Leu	Asn	Ser	Gln	Thr	Val	Lys	His	Leu	
		35					40					45				
Asp	Thr	Leu	His	Glu	Tyr	Leu	Gly	Met	Met	Met	Ala	Val	Gln	Gly	Ala	
	50					55					60					
Phe	Ala	Asp	Arg	Ser	Ser	Ala	Leu	Leu	Thr	Val	Gln	Thr	Leu	Leu	Ser	
65					70					75					80	
Glu	Leu	Pro	Ser	Leu	Gln	Thr	Arg	Val	Glu	Lys	Leu	Glu	Ala	Ala	Ser	
			85					90						95		
Ser	Lys	Val	Phe	Gly	Gly	Asp	Lys	Ser	Arg	Ile	Arg	Lys	Ile	Glu	Glu	
		100						105					110			
Leu	Lys	Glu	Thr	Ile	Lys	Val	Thr	Glu	Asp	Ala	Lys	Asn	Val	Ala	Ile	
		115					120					125				
Lys	Gly	Tyr	Glu	Arg	Ile	Lys	Glu	Asn	Asn	Arg	Ser	Glu	Val	Glu	Arg	
	130					135					140					
Leu	Asp	Arg	Glu	Arg	Arg	Ala	Asp	Phe	Met	Asn	Met	Met	Lys	Gly	Phe	
145					150					155					160	
Val	Val	Asn	Gln	Val	Gly	Tyr	Ala	Glu	Lys	Met	Gly	Asn	Val	Trp	Ala	
			165					170						175		
Lys	Val	Ala	Glu	Thr	Ser	Gln	Tyr	Asp	Arg	Glu	Lys	Gln	Ser	Ser		
		180						185						190		

(2) INFORMATION FOR SEQ ID NO:54:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 170 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..170

(D) OTHER INFORMATION: / Ceres Seq. ID 1481522

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:54:

Met	Lys	Asn	Leu	Ala	Thr	Ala	Ala	Val	Lys	Ala	Ser	Arg	Phe	Tyr	Arg	
1			5					10					15			
Glu	Leu	Asn	Ser	Gln	Thr	Val	Lys	His	Leu	Asp	Thr	Leu	His	Glu	Tyr	
		20					25					30				
Leu	Gly	Met	Met	Met	Ala	Val	Gln	Gly	Ala	Phe	Ala	Asp	Arg	Ser	Ser	
	35					40					45					
Ala	Leu	Leu	Thr	Val	Gln	Thr	Leu	Leu	Ser	Glu	Leu	Pro	Ser	Leu	Gln	
	50					55				60						
Thr	Arg	Val	Glu	Lys	Leu	Glu	Ala	Ala	Ser	Ser	Lys	Val	Phe	Gly	Gly	
65				70					75						80	
Asp	Lys	Ser	Arg	Ile	Arg	Lys	Ile	Glu	Glu	Leu	Lys	Glu	Thr	Ile	Lys	
		85					90						95			
Val	Thr	Glu	Asp	Ala	Lys	Asn	Val	Ala	Ile	Lys	Gly	Tyr	Glu	Arg	Ile	
		100					105						110			
Lys	Glu	Asn	Asn	Arg	Ser	Glu	Val	Glu	Arg	Leu	Asp	Arg	Glu	Arg	Arg	
	115						120						125			
Ala	Asp	Phe	Met	Asn	Met	Met	Lys	Gly	Phe	Val	Val	Asn	Gln	Val	Gly	
	130					135					140					
Tyr	Ala	Glu	Lys	Met	Gly	Asn	Val	Trp	Ala	Lys	Val	Ala	Glu	Glu	Thr	
145				150					155						160	
Ser	Gln	Tyr	Asp	Arg	Glu	Lys	Gln	Ser	Ser							
		165						170								

(2) INFORMATION FOR SEQ ID NO:55:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 136 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..136
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481523

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:55:

```
Met Met Met Ala Val Gln Gly Ala Phe Ala Asp Arg Ser Ser Ala Leu
1          5          10          15
Leu Thr Val Gln Thr Leu Leu Ser Glu Leu Pro Ser Leu Gln Thr Arg
20          25          30
Val Glu Lys Leu Glu Ala Ala Ser Ser Lys Val Phe Gly Gly Asp Lys
35          40          45
Ser Arg Ile Arg Lys Ile Glu Glu Leu Lys Glu Thr Ile Lys Val Thr
50          55          60
Glu Asp Ala Lys Asn Val Ala Ile Lys Gly Tyr Glu Arg Ile Lys Glu
65          70          75          80
Asn Asn Arg Ser Glu Val Glu Arg Leu Asp Arg Glu Arg Arg Ala Asp
85          90          95
Phe Met Asn Met Met Lys Gly Phe Val Val Asn Gln Val Gly Tyr Ala
100          105          110
Glu Lys Met Gly Asn Val Trp Ala Lys Val Ala Glu Glu Thr Ser Gln
115          120          125
Tyr Asp Arg Glu Lys Gln Ser Ser
130          135
```

(2) INFORMATION FOR SEQ ID NO:56:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 2180 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..2180
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481524

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:56:

```
attactcaat tagtacaagt tggtatacaa ctaaattcttc atttggtataa tcattctttac      60
tcaaattgaa tagtagtggt cgtgtgaaaa caagaaaagt ggaaaaggac aaaagagaga      120
gtaaaggacg cctcctaata aagcactctt cttccttttc actttcctca ttgaagagag      180
agccaaattc agcttaaagc cccataagcg taagcgtaag cgtaagcgta agcgtaagcg      240
taagcgtaag cgtaagcgta agcgtaagcg taagcgtaag cgcggggata aatctctctc      300
ttcctcacct gcgttttcgt ggagcatctt cttcaacaat ggctgcttct ccgatctgat      360
catatcctga tttgaatttt gctatctctc atgcctcgaa ctcgttttgt cgacgtagca      420
tcctagtgcg tgaggaagaa gaagaagatg agcttcttta tcctctccgt cgtcgttttc      480
gtttctctcg ctttcttctc tcttcgcgat tccgttgatt catctgtttc cgcttcacag      540
gatactctca gactcatatt aggttcaccg aattttggaa catggaaagg tggaaatctca      600
ttagcaccag gacctgaatc tgatgatgtt gtctctgatt acctcctctt agcagctcat      660
agaaccaaga gacctgacat tcttagagct tttaagcctt accatgggtg ctggaacatc      720
accaataatc actattgggc ttctgttgga ttacaggtg ctccctgggtt catactagct      780
gttatctggc tcttgtcttt tggctctctt cttgttggtt atcattgctt caaatggaga      840
atatgtgata aagctaaagg atcatcattc gatacacgaa gaatctgttt cattttgttg      900
attgtgttta catgtgttg agcggtggga tgcattcttt tatctgttg acaagataag      960
tttcataccg aagctatgca tactcttaag tatgttgtaa accagtcaga ctacactgtg      1020
```

```
gagatcctcc agaatgtgac tcaatatctg tcccttgcca aaacgattaa cgtgacacag 1080
attgtcattc cgtctgatgt aatgggtgaa attgacaagt taaatgtcaa tcttaacact 1140
gcagctgtaa cactgggaga gacaacaaca gataaacgct gctaaaatta agagagtttt 1200
ctatgctgtg cgatcagctt tgatcacggt cgctactgtg atgctcatcc tttcttttgt 1260
aggtctattg ctttctgtcc tccgccacca acatgttggt catatatcgc tcgtgagtgg 1320
gtggatactt gtggctgtga catttggtct ttgtggagtc tttctgatcc taaacaatgc 1380
aatttctgat acgtgtgtag caatgaagga atgggttgat aatcctcacg cagaaacagc 1440
tctcagcagc attctcccat gcgttgatca gcaacaaca aaccagactc tttcacagag 1500
taaagtgtgt atcaacagca tcgtgaccgt tgtaaaccacc tttgtctatg ctgttgccaa 1560
tacaaaccca gctccaggtc aagaccgcta ttacaaccag tctggacctc cgatgcctcc 1620
tttatgcatc ccatttgatg caaacatgga agatcgccag tgctcgctt gggaactatc 1680
aatagaaaat gcatcatcgg tctgggagaa ttacaaatgc gaggttacac catctggaat 1740
ctgcaccacc gtggggagag taacgccaga tacctttgga cagttggtag cagctgtgaa 1800
tgagagctac gctctagagc attacacgcc tccattgctt agcttccgag attgcaactt 1860
tgttagggaa acatttatga gtattacctc agattactgt ccaccgttag tacgtaatct 1920
gaggattgtg aacgcaggac tccgactgat ctccgtagga gtcttactat gtctggtgct 1980
atggatatct tatgcgaacc cccccaagg gaggaagtgt ttgcggatcc acaccctcaa 2040
agaaaagatg atagctttgg taacggcttg gatactcatc actcagatga cgaacctaa 2100
ctttctgtag aatgcgtata gtatagaggt atagttagat agaatcagat attgtatttc 2160
ataatgatat gaaagagaac
```

(2) INFORMATION FOR SEQ ID NO:57:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 245 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..245
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481525

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:57:

```
Met Ser Phe Phe Ile Leu Ser Val Val Val Phe Val Ser Leu Ala Phe
1           5           10           15
Phe Ser Leu Pro His Ser Val Asp Ser Ser Val Ser Ala Ser Gln Asp
           20           25           30
Pro Leu Arg Leu Ile Leu Gly Ser Pro Asn Phe Gly Thr Trp Lys Gly
           35           40           45
Gly Ile Ser Leu Ala Pro Gly Pro Glu Ser Asp Asp Val Val Ser Asp
           50           55           60
Tyr Leu Leu Leu Ala Ala His Arg Thr Lys Arg Pro Asp Ile Leu Arg
65           70           75           80
Ala Phe Lys Pro Tyr His Gly Gly Trp Asn Ile Thr Asn Asn His Tyr
           85           90           95
Trp Ala Ser Val Gly Phe Thr Gly Ala Pro Gly Phe Ile Leu Ala Val
           100          105          110
Ile Trp Leu Leu Ser Phe Gly Ser Leu Leu Val Val Tyr His Cys Phe
           115          120          125
Lys Trp Arg Ile Cys Asp Lys Ala Lys Gly Ser Ser Phe Asp Thr Arg
           130          135          140
Arg Ile Cys Phe Ile Leu Leu Ile Val Phe Thr Cys Val Ala Ala Val
145          150          155          160
Gly Cys Ile Leu Leu Ser Val Gly Gln Asp Lys Phe His Thr Glu Ala
           165          170          175
Met His Thr Leu Lys Tyr Val Val Asn Gln Ser Asp Tyr Thr Val Glu
           180          185          190
Ile Leu Gln Asn Val Thr Gln Tyr Leu Leu Ala Lys Thr Ile Asn
           195          200          205
Val Thr Gln Ile Val Ile Pro Ser Asp Val Met Gly Glu Ile Asp Lys
```

210 215 220
Leu Asn Val Asn Leu Asn Thr Ala Ala Val Thr Leu Gly Glu Thr Thr
225 230 235 240
Thr Asp Lys Arg Cys
245

(2) INFORMATION FOR SEQ ID NO:58:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 289 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..289
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481526

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:58:

Met Leu Ile Leu Ser Phe Val Gly Leu Leu Leu Ser Val Leu Arg His
1 5 10 15
Gln His Val Val His Ile Phe Val Val Ser Gly Trp Ile Leu Val Ala
20 25 30
Val Thr Phe Val Leu Cys Gly Val Phe Leu Ile Leu Asn Ala Ile
35 40 45
Ser Asp Thr Cys Val Ala Met Lys Glu Trp Val Asp Asn Pro His Ala
50 55 60
Glu Thr Ala Leu Ser Ser Ile Leu Pro Cys Val Asp Gln Gln Thr Thr
65 70 75 80
Asn Gln Thr Leu Ser Gln Ser Lys Val Val Ile Asn Ser Ile Val Thr
85 90 95
Val Val Asn Thr Phe Val Tyr Ala Val Ala Asn Thr Asn Pro Ala Pro
100 105 110
Gly Gln Asp Arg Tyr Tyr Asn Gln Ser Gly Pro Pro Met Pro Pro Leu
115 120 125
Cys Ile Pro Phe Asp Ala Asn Met Glu Asp Arg Gln Cys Ser Pro Trp
130 135 140
Glu Leu Ser Ile Glu Asn Ala Ser Ser Val Trp Glu Asn Tyr Lys Cys
145 150 155 160
Glu Val Thr Pro Ser Gly Ile Cys Thr Thr Val Gly Arg Val Thr Pro
165 170 175
Asp Thr Phe Gly Gln Leu Val Ala Ala Val Asn Glu Ser Tyr Ala Leu
180 185 190
Glu His Tyr Thr Pro Pro Leu Leu Ser Phe Arg Asp Cys Asn Phe Val
195 200 205
Arg Glu Thr Phe Met Ser Ile Thr Ser Asp Tyr Cys Pro Pro Leu Val
210 215 220
Arg Asn Leu Arg Ile Val Asn Ala Gly Leu Gly Leu Ile Ser Val Gly
225 230 235 240
Val Leu Leu Cys Leu Val Leu Trp Ile Phe Tyr Ala Asn Pro Pro Lys
245 250 255
Gly Arg Lys Cys Leu Arg Ile His Thr Leu Lys Glu Lys Met Ile Ala
260 265 270
Leu Val Thr Ala Trp Ile Leu Ile Thr Gln Met Thr Asn Leu Ser Phe
275 280 285
Leu

(2) INFORMATION FOR SEQ ID NO:59:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 235 amino acids
- (B) TYPE: amino acid

(C) STRANDEDNESS:
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..235
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481527
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:59:
Met Lys Glu Trp Val Asp Asn Pro His Ala Glu Thr Ala Leu Ser Ser
1 5 10 15
Ile Leu Pro Cys Val Asp Gln Gln Thr Thr Asn Gln Thr Leu Ser Gln
 20 25 30
Ser Lys Val Val Ile Asn Ser Ile Val Thr Val Val Asn Thr Phe Val
 35 40 45
Tyr Ala Val Ala Asn Thr Asn Pro Ala Pro Gly Gln Asp Arg Tyr Tyr
 50 55 60
Asn Gln Ser Gly Pro Pro Met Pro Pro Leu Cys Ile Pro Phe Asp Ala
65 70 75 80
Asn Met Glu Asp Arg Gln Cys Ser Pro Trp Glu Leu Ser Ile Glu Asn
 85 90 95
Ala Ser Ser Val Trp Glu Asn Tyr Lys Cys Glu Val Thr Pro Ser Gly
 100 105 110
Ile Cys Thr Thr Val Gly Arg Val Thr Pro Asp Thr Phe Gly Gln Leu
 115 120 125
Val Ala Ala Val Asn Glu Ser Tyr Ala Leu Glu His Tyr Thr Pro Pro
 130 135 140
Leu Leu Ser Phe Arg Asp Cys Asn Phe Val Arg Glu Thr Phe Met Ser
145 150 155 160
Ile Thr Ser Asp Tyr Cys Pro Pro Leu Val Arg Asn Leu Arg Ile Val
 165 170 175
Asn Ala Gly Leu Gly Leu Ile Ser Val Gly Val Leu Leu Cys Leu Val
 180 185 190
Leu Trp Ile Phe Tyr Ala Asn Pro Pro Lys Gly Arg Lys Cys Leu Arg
 195 200 205
Ile His Thr Leu Lys Glu Lys Met Ile Ala Leu Val Thr Ala Trp Ile
210 215 220
Leu Ile Thr Gln Met Thr Asn Leu Ser Phe Leu
225 230 235

(2) INFORMATION FOR SEQ ID NO:60:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 634 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: DNA (genomic)
 (ix) FEATURE:

 (A) NAME/KEY: -
 (B) LOCATION: 1..634
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481532
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:60:
aaaacatctc tcgccgtcag gttacatcta tcgccaccgc aaagagacca ccgtctcctc 60
cgcaatcttc ataacctaaa caacctcat cccctggtac ttaaacaatg ggaaagagga 120
aatcaagagc aaagcctgct cctacgaagc gaatggataa gcttgacaca atctttagtt 180
gtcctttctg caatcacggg tctagtgtcg aatgcatcat tgatatgaag catctgattg 240
gtaaagcagc ttgtagaatc tgtgaagaaa gctttaggta ctactatcac agctttgact 300
gaagctatag acatttatag cgaatggatc gatgagtgcg agagggttaa taccgcggaa 360
gatgatgttg tgcaagaaga ggaggatgat gaagatgacc atgtctctgt caaaaggaag 420
tataacttct gagacgagtg ttttatcgaa aatcatgtaa gtcgtcgtct tagagttatc 480
tgctttatgt tgtaatatct atctgatgaa atcacaagaa caatctttag tgttttctca 540

gtgtctgata gagaacata catttaagtg aacaatcttt aatcacaata acagtgtatg 600
attatgattt gtaagtggat ttaaggcttt gctt

(2) INFORMATION FOR SEQ ID NO:61:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 75 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..75
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481533

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:61:

Lys	Thr	Ser	Leu	Ala	Val	Arg	Leu	His	Leu	Ser	Pro	Pro	Gln	Arg	Asp
1			5					10					15		
His	Arg	Leu	Leu	Arg	Asn	Leu	His	Asn	Leu	Asn	Asn	Pro	His	Pro	Leu
		20					25					30			
Val	Leu	Lys	Gln	Trp	Glu	Arg	Gly	Asn	Gln	Glu	Gln	Ser	Leu	Leu	Leu
		35				40					45				
Arg	Ser	Glu	Trp	Ile	Ser	Leu	Thr	Gln	Ser	Leu	Val	Val	Leu	Ser	Ala
	50				55					60					
Ile	Thr	Gly	Leu	Val	Ser	Asn	Ala	Ser	Leu	Ile					
65			70						75						

(2) INFORMATION FOR SEQ ID NO:62:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 64 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..64
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481534

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:62:

Met	Gly	Lys	Arg	Lys	Ser	Arg	Ala	Lys	Pro	Ala	Pro	Thr	Lys	Arg	Met
1			5					10					15		
Asp	Lys	Leu	Asp	Thr	Ile	Phe	Ser	Cys	Pro	Phe	Cys	Asn	His	Gly	Ser
		20					25					30			
Ser	Val	Glu	Cys	Ile	Ile	Asp	Met	Lys	His	Leu	Ile	Gly	Lys	Ala	Ala
		35				40					45				
Cys	Arg	Ile	Cys	Glu	Glu	Ser	Phe	Arg	Tyr	Tyr	Tyr	His	Ser	Phe	Asp
	50					55					60				

(2) INFORMATION FOR SEQ ID NO:63:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 49 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..49
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481535

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:63:

Met Asp Lys Leu Asp Thr Ile Phe Ser Cys Pro Phe Cys Asn His Gly

1	5	10	15
Ser Ser Val Glu Cys Ile Ile Asp Met Lys His Leu Ile Gly Lys Ala			
	20	25	30
Ala Cys Arg Ile Cys Glu Glu Ser Phe Arg Tyr Tyr Tyr His Ser Phe			
	35	40	45

Asp

(2) INFORMATION FOR SEQ ID NO:64:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1668 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1668
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481540

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:64:

```
atctgcattg ttctccgcct ctctctctca aactcttcag tttgcaaaac ccttaagaag      60
gtgtgaatta gtaagtaatg gggaagaaga agttttattga taagaaaaaag gcggcgactt      120
tcgagttgtg tcctcgtgat acgtcagacc caagatacag tgatgcacca ggtggtgata      180
agatcttctt acgagttgat caaaaccctg ttaacatcaa tggtttcatt gaagaagatg      240
aagaagattt tagagttagg gtatcctgat gatggttata attacttgga gcatttgaga      300
gagattaaga atactgggtg tggttctaata ttctatgtga atcctaagta tgaggttgct      360
cagttacctc gtgatgttaa ggcttatgat gcgtctcgtg ttaagatctc tggtatgggtg      420
aatgaagaag gtaatgataa taagttgatg tatagtgttg cgtccaagac tgttaacgtc      480
aaggtgcaga aagctattga tcctgaagtt gctgcgttgc ttgaaaacag tgatgggtct      540
gagtttggtt ctgatgttga ggatttgga gaagattttg ttggttcaagc taatcttact      600
caaaaggggtg aatcttcttg tgtgagcaat ggagagctcg agttttctgt aagacgtgag      660
gttagagaaa gagaaagtga tgaacctgtg gctgaaaacc cgagagttcc tcgtcaaatt      720
gatgagctat ttgatcagct cgaactcaat gaatatggaa gtgatagtga cggatgatgg      780
tacatagctg aagatggaga agaagaagaa gaagaagact tcatggctca agaagttcag      840
aatcttattc atgggaaggc aaaagattat gagcttgaag aaaaatataat gaaccctgcg      900
gatatactga agaacagtga ctctgtcaga gataaagagg aagtggacac tgctgctcat      960
gttatccgcc gaactgtaga atatggtgaa aattttgata acgggaatga agatgaattt     1020
gtagagctga ctgaagaaaag cagcgatgaa agcgagaagc atgattgtga aaccatagtc     1080
tcaacatact cgaatctcga taacctccct ggtaaaatcc ttgctgcaga gtcagctagg     1140
cagaagaagc tgagtgaaac attagctaac gcattgagtt caaatggaag aatcattaat     1200
ctccaaggga gagagaggat tcctgtcga tttttacctg gtaggagagc tgaacaaacc     1260
gatgtcaaag cggaaatccc aaaagctgaa ccgatcaaga ggaagactca tgggtcaagag     1320
tcgaaagaag agaagaaaaga gcgggaaaaat gctgtaaaag ccgaaaagcg agaagcaagg     1380
ataattaaga aacagacaaa gatgctgtat tgcggtgaaa cgcagcgtgc tcaaagagct     1440
gttgctacct ctggtccatc gtcgagacct ctaaaataat atgttactaa ggtaaaaacaa     1500
aacaattctc agactgttta aaaccagttt ttccagccat ttcgtgtaat atttgctgtt     1560
tgtttttttc tttttcatca agatttgaaa atcttgaatc ttgttttgga tgtggacgtt     1620
ttgaatatta tttattactt ttactagtct aatttcgaga aagtgatg
```

(2) INFORMATION FOR SEQ ID NO:65:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 438 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..438
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481541

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:65:

Met	His	Gln	Val	Val	Ile	Arg	Ser	Ser	Tyr	Glu	Leu	Ile	Lys	Thr	Leu
1				5					10					15	
Leu	Thr	Ser	Met	Val	Ser	Leu	Lys	Lys	Met	Lys	Lys	Ile	Leu	Glu	Leu
		20						25					30		
Gly	Tyr	Pro	Asp	Asp	Gly	Tyr	Asn	Tyr	Leu	Glu	His	Leu	Arg	Glu	Ile
		35					40					45			
Lys	Asn	Thr	Gly	Gly	Gly	Ser	Asn	Phe	Tyr	Val	Asn	Pro	Lys	Tyr	Glu
	50					55					60				
Val	Ala	Gln	Leu	Pro	Arg	Asp	Val	Lys	Ala	Tyr	Asp	Ala	Ser	Arg	Val
65					70					75				80	
Lys	Ile	Ser	Gly	Met	Val	Asn	Glu	Glu	Gly	Asn	Asp	Asn	Lys	Leu	Met
				85					90					95	
Tyr	Ser	Val	Ala	Ser	Lys	Thr	Val	Asn	Val	Lys	Val	Gln	Lys	Ala	Ile
			100					105					110		
Asp	Pro	Glu	Val	Ala	Ala	Leu	Leu	Glu	Asn	Ser	Asp	Gly	Ser	Glu	Phe
		115						120				125			
Gly	Ser	Asp	Val	Glu	Asp	Leu	Glu	Glu	Asp	Phe	Val	Val	Gln	Ala	Asn
	130					135					140				
Leu	Thr	Gln	Lys	Gly	Glu	Ser	Ser	Gly	Val	Ser	Asn	Gly	Glu	Leu	Glu
145					150					155				160	
Phe	Ser	Val	Arg	Arg	Glu	Val	Arg	Glu	Arg	Glu	Ser	Asp	Glu	Pro	Val
				165					170					175	
Ala	Glu	Asn	Pro	Arg	Val	Pro	Arg	Gln	Ile	Asp	Glu	Leu	Phe	Asp	Gln
			180					185					190		
Leu	Glu	Leu	Asn	Glu	Tyr	Gly	Ser	Asp	Ser	Asp	Gly	Asp	Gly	Tyr	Ile
		195					200					205			
Ala	Glu	Asp	Gly	Glu	Glu	Glu	Glu	Glu	Glu	Asp	Phe	Met	Ala	Gln	Glu
	210					215					220				
Val	Gln	Asn	Leu	Ile	His	Gly	Lys	Ala	Lys	Asp	Tyr	Glu	Leu	Glu	Glu
225					230					235				240	
Lys	Tyr	Met	Asn	Pro	Ala	Asp	Ile	Leu	Lys	Asn	Ser	Asp	Ser	Val	Arg
			245						250					255	
Asp	Lys	Glu	Glu	Val	Asp	Thr	Ala	Ala	His	Val	Ile	Arg	Arg	Thr	Val
			260					265					270		
Glu	Tyr	Gly	Glu	Asn	Phe	Asp	Asn	Gly	Asn	Glu	Asp	Glu	Phe	Val	Glu
	275						280					285			
Leu	Thr	Glu	Glu	Ser	Ser	Asp	Glu	Ser	Glu	Lys	His	Asp	Cys	Glu	Thr
290						295					300				
Ile	Val	Ser	Thr	Tyr	Ser	Asn	Leu	Asp	Asn	Leu	Pro	Gly	Lys	Ile	Leu
305					310					315				320	
Ala	Ala	Glu	Ser	Ala	Arg	Gln	Lys	Lys	Leu	Ser	Glu	Thr	Leu	Ala	Asn
				325					330					335	
Ala	Leu	Ser	Ser	Asn	Gly	Arg	Ile	Ile	Asn	Leu	Gln	Gly	Arg	Glu	Arg
			340					345					350		
Ile	Pro	Val	Glu	Phe	Leu	Pro	Gly	Arg	Arg	Ala	Glu	Gln	Thr	Asp	Val
		355					360						365		
Lys	Ala	Glu	Ile	Pro	Lys	Ala	Glu	Pro	Ile	Lys	Arg	Lys	Thr	His	Gly
	370					375					380				
Gln	Glu	Ser	Lys	Glu	Glu	Lys	Lys	Glu	Arg	Lys	Asn	Ala	Val	Lys	Ala
385					390					395				400	
Glu	Lys	Arg	Glu	Ala	Arg	Ile	Ile	Lys	Lys	Gln	Thr	Lys	Met	Leu	Tyr
				405					410					415	
Cys	Gly	Glu	Thr	Gln	Arg	Ala	Gln	Arg	Ala	Val	Ala	Thr	Ser	Gly	Pro
			420					425					430		
Ser	Ser	Arg	Pro	Leu	Lys										

(2) INFORMATION FOR SEQ ID NO:66:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 419 amino acids

(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..419
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481542
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:66:

Met	Val	Ser	Leu	Lys	Lys	Met	Lys	Lys	Ile	Leu	Glu	Leu	Gly	Tyr	Pro
1				5				10						15	
Asp	Asp	Gly	Tyr	Asn	Tyr	Leu	Glu	His	Leu	Arg	Glu	Ile	Lys	Asn	Thr
			20					25					30		
Gly	Gly	Gly	Ser	Asn	Phe	Tyr	Val	Asn	Pro	Lys	Tyr	Glu	Val	Ala	Gln
		35					40					45			
Leu	Pro	Arg	Asp	Val	Lys	Ala	Tyr	Asp	Ala	Ser	Arg	Val	Lys	Ile	Ser
	50					55					60				
Gly	Met	Val	Asn	Glu	Glu	Gly	Asn	Asp	Asn	Lys	Leu	Met	Tyr	Ser	Val
65				70					75					80	
Ala	Ser	Lys	Thr	Val	Asn	Val	Lys	Val	Gln	Lys	Ala	Ile	Asp	Pro	Glu
				85					90					95	
Val	Ala	Ala	Leu	Leu	Glu	Asn	Ser	Asp	Gly	Ser	Glu	Phe	Gly	Ser	Asp
			100					105					110		
Val	Glu	Asp	Leu	Glu	Glu	Asp	Phe	Val	Val	Gln	Ala	Asn	Leu	Thr	Gln
		115						120				125			
Lys	Gly	Glu	Ser	Ser	Gly	Val	Ser	Asn	Gly	Glu	Leu	Glu	Phe	Ser	Val
	130					135					140				
Arg	Arg	Glu	Val	Arg	Glu	Arg	Glu	Ser	Asp	Glu	Pro	Val	Ala	Glu	Asn
145					150					155				160	
Pro	Arg	Val	Pro	Arg	Gln	Ile	Asp	Glu	Leu	Phe	Asp	Gln	Leu	Glu	Leu
				165				170						175	
Asn	Glu	Tyr	Gly	Ser	Asp	Ser	Asp	Gly	Asp	Gly	Tyr	Ile	Ala	Glu	Asp
		180						185					190		
Gly	Glu	Glu	Glu	Glu	Glu	Glu	Asp	Phe	Met	Ala	Gln	Glu	Val	Gln	Asn
		195					200					205			
Leu	Ile	His	Gly	Lys	Ala	Lys	Asp	Tyr	Glu	Leu	Glu	Glu	Lys	Tyr	Met
	210					215					220				
Asn	Pro	Ala	Asp	Ile	Leu	Lys	Asn	Ser	Asp	Ser	Val	Arg	Asp	Lys	Glu
225					230					235				240	
Glu	Val	Asp	Thr	Ala	Ala	His	Val	Ile	Arg	Thr	Val	Glu	Tyr	Gly	
			245						250					255	
Glu	Asn	Phe	Asp	Asn	Gly	Asn	Glu	Asp	Glu	Phe	Val	Glu	Leu	Thr	Glu
		260						265					270		
Glu	Ser	Ser	Asp	Glu	Ser	Glu	Lys	His	Asp	Cys	Glu	Thr	Ile	Val	Ser
	275						280					285			
Thr	Tyr	Ser	Asn	Leu	Asp	Asn	Leu	Pro	Gly	Lys	Ile	Leu	Ala	Ala	Glu
	290					295					300				
Ser	Ala	Arg	Gln	Lys	Lys	Leu	Ser	Glu	Thr	Leu	Ala	Asn	Ala	Leu	Ser
305				310						315				320	
Ser	Asn	Gly	Arg	Ile	Asn	Leu	Gln	Gly	Arg	Glu	Arg	Ile	Pro	Val	
			325					330					335		
Glu	Phe	Leu	Pro	Gly	Arg	Arg	Ala	Glu	Gln	Thr	Asp	Val	Lys	Ala	Glu
		340					345						350		
Ile	Pro	Lys	Ala	Glu	Pro	Ile	Lys	Arg	Lys	Thr	His	Gly	Gln	Glu	Ser
	355					360						365			
Lys	Glu	Glu	Lys	Lys	Glu	Arg	Lys	Asn	Ala	Val	Lys	Ala	Glu	Lys	Arg
	370					375					380				
Glu	Ala	Arg	Ile	Ile	Lys	Lys	Gln	Thr	Lys	Met	Leu	Tyr	Cys	Gly	Glu
385				390						395					400

Thr Gln Arg Ala Gln Arg Ala Val Ala Thr Ser Gly Pro Ser Ser Arg
405 410 415
Pro Leu Lys

(2) INFORMATION FOR SEQ ID NO:67:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 413 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..413
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481543

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:67:

Met	Lys	Lys	Ile	Leu	Glu	Leu	Gly	Tyr	Pro	Asp	Asp	Gly	Tyr	Asn	Tyr
1			5					10						15	
Leu	Glu	His	Leu	Arg	Glu	Ile	Lys	Asn	Thr	Gly	Gly	Gly	Ser	Asn	Phe
		20					25						30		
Tyr	Val	Asn	Pro	Lys	Tyr	Glu	Val	Ala	Gln	Leu	Pro	Arg	Asp	Val	Lys
		35				40						45			
Ala	Tyr	Asp	Ala	Ser	Arg	Val	Lys	Ile	Ser	Gly	Met	Val	Asn	Glu	Glu
	50					55					60				
Gly	Asn	Asp	Asn	Lys	Leu	Met	Tyr	Ser	Val	Ala	Ser	Lys	Thr	Val	Asn
65				70						75				80	
Val	Lys	Val	Gln	Lys	Ala	Ile	Asp	Pro	Glu	Val	Ala	Ala	Leu	Leu	Glu
			85					90					95		
Asn	Ser	Asp	Gly	Ser	Glu	Phe	Gly	Ser	Asp	Val	Glu	Asp	Leu	Glu	Glu
		100					105						110		
Asp	Phe	Val	Val	Gln	Ala	Asn	Leu	Thr	Gln	Lys	Gly	Glu	Ser	Ser	Gly
		115					120					125			
Val	Ser	Asn	Gly	Glu	Leu	Glu	Phe	Ser	Val	Arg	Arg	Glu	Val	Arg	Glu
	130					135					140				
Arg	Glu	Ser	Asp	Glu	Pro	Val	Ala	Glu	Asn	Pro	Arg	Val	Pro	Arg	Gln
145				150						155				160	
Ile	Asp	Glu	Leu	Phe	Asp	Gln	Leu	Glu	Leu	Asn	Glu	Tyr	Gly	Ser	Asp
			165					170						175	
Ser	Asp	Gly	Asp	Gly	Tyr	Ile	Ala	Glu	Asp	Gly	Glu	Glu	Glu	Glu	Glu
		180					185						190		
Glu	Asp	Phe	Met	Ala	Gln	Glu	Val	Gln	Asn	Leu	Ile	His	Gly	Lys	Ala
	195					200					205				
Lys	Asp	Tyr	Glu	Leu	Glu	Glu	Lys	Tyr	Met	Asn	Pro	Ala	Asp	Ile	Leu
	210					215					220				
Lys	Asn	Ser	Asp	Ser	Val	Arg	Asp	Lys	Glu	Glu	Val	Asp	Thr	Ala	Ala
225				230						235				240	
His	Val	Ile	Arg	Arg	Thr	Val	Glu	Tyr	Gly	Glu	Asn	Phe	Asp	Asn	Gly
			245					250						255	
Asn	Glu	Asp	Glu	Phe	Val	Glu	Leu	Thr	Glu	Glu	Ser	Ser	Asp	Glu	Ser
		260					265						270		
Glu	Lys	His	Asp	Cys	Glu	Thr	Ile	Val	Ser	Thr	Tyr	Ser	Asn	Leu	Asp
	275					280					285				
Asn	Leu	Pro	Gly	Lys	Ile	Leu	Ala	Ala	Glu	Ser	Ala	Arg	Gln	Lys	Lys
	290					295					300				
Leu	Ser	Glu	Thr	Leu	Ala	Asn	Ala	Leu	Ser	Ser	Asn	Gly	Arg	Ile	Ile
305				310						315				320	
Asn	Leu	Gln	Gly	Arg	Glu	Arg	Ile	Pro	Val	Glu	Phe	Leu	Pro	Gly	Arg
			325					330					335		
Arg	Ala	Glu	Gln	Thr	Asp	Val	Lys	Ala	Glu	Ile	Pro	Lys	Ala	Glu	Pro

```

          340          345          350
Ile Lys Arg Lys Thr His Gly Gln Glu Ser Lys Glu Glu Lys Lys Glu
          355          360          365
Arg Lys Asn Ala Val Lys Ala Glu Lys Arg Glu Ala Arg Ile Ile Lys
          370          375          380
Lys Gln Thr Lys Met Leu Tyr Cys Gly Glu Thr Gln Arg Ala Gln Arg
385          390          395          400
Ala Val Ala Thr Ser Gly Pro Ser Ser Arg Pro Leu Lys
          405          410
```

(2) INFORMATION FOR SEQ ID NO:68:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1601 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1601
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481544

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:68:

```

aatcttcttc attcgaaggt tactcgcaact tcctctgcac acttcttccct ctccatctaa      60
cctcagtact cctacaatcc tctaagaatc catagatcta ctgctgggaa aagcttcgcg      120
acaatgtctt ggcctacgga ttctgagtta aattccataa aggaggcagt ggctcagatg      180
agtggaagag ataaaggaga agttcgagtg gtggtcgctc cttatcgatat atgtccttta      240
ggagctcaca ttgatcacca ggggtggaact gtatcagcta tgacgattaa agggatcctt      300
cttggtttttg ttccatcggg tgatactcag gtccagttgc gctctgcaca atttgaagga      360
gaagtatgtt tcagagtaga tgaaatccag caccaatag gcctagcaaa caagaatggg      420
gcaagcacgc catctccatc gaaggaaaaa agtatctggg gtacttatgc cagaggagca      480
gtttatgcgt tacagagcag caaaaagaat ctcaaacagg gcattattgg ttacctcagt      540
ggctcaaattg gactagatag ctccgggctt agctcatcag ctgctgttgg tgtggcatac      600
ctgctagctc tagagaatgc aaacgaattg actgtatccc caacagaaaa tatcgaatat      660
gacaggctta ttgagaatgt gtatctgggt ctgcggaatg gaattttgga tcaatcagct      720
attttgcttt cgaattatgg gtgtctaaca tacatggact gcaagacttt ggaccacgag      780
cttgtacagg ctctgaact ggagaaaccg ttcaggatat tgttagcatt ctcaggcttg      840
aggcaggcgt tgaccaccaa cccaggatat aatctgcgag tttctgagtg tcaagaggca      900
gcaaaagttc ttttgactgc atctgggaac agtgagctgg aacctacgtt gtgcaatggt      960
gagcatgcgg tctatgaagc tcacaagcat gagctgaaac cggtttttagc taaaagagca     1020
gagcattatt tctcggagaa catgcgagtt atcaaaggac gggaagcctg ggcttcaggg     1080
aatcttgaag aatttggaag gctaatttca gcatccggct tgagttccat tgagaattac     1140
gaatgcgggt cggagccact gatccagcta tacaagattc ttctgaaggc tcctgggtga     1200
tatggagcta gattcagcgg tgcaggtttc aggggatggt gtctagcctt tgtagatgca     1260
gtaaaagctg aggcagctgc ttcatatgtg aaggatgaat atgaaaaggc ccaaccgag     1320
tttgctaaca atctaaatgg aggaaaacct gttctcatct gtgaagcagg tgacgctgct     1380
cgtgttcttc tctgatcaat cctggagttt ttggtttctt cccacttaa actcgatttt     1440
tttgtccctt atatctctca cgcttattga ttctttgctt gtttatctct ttttgatcct     1500
gtctgagaaa ttctctggtc tctttggctg gagtttcatac attgcttgat acattttttt     1560
tgctacaaat acataatgta aatcattctc taccgttttc c
```

(2) INFORMATION FOR SEQ ID NO:69:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 423 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..423
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481545

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:69:

Met	Ser	Trp	Pro	Thr	Asp	Ser	Glu	Leu	Asn	Ser	Ile	Lys	Glu	Ala	Val
1				5				10						15	
Ala	Gln	Met	Ser	Gly	Arg	Asp	Lys	Gly	Glu	Val	Arg	Val	Val	Val	Ala
		20						25					30		
Pro	Tyr	Arg	Ile	Cys	Pro	Leu	Gly	Ala	His	Ile	Asp	His	Gln	Gly	Gly
		35					40					45			
Thr	Val	Ser	Ala	Met	Thr	Ile	Lys	Gly	Ile	Leu	Leu	Gly	Phe	Val	Pro
	50					55					60				
Ser	Gly	Asp	Thr	Gln	Val	Gln	Leu	Arg	Ser	Ala	Gln	Phe	Glu	Gly	Glu
65					70					75					80
Val	Cys	Phe	Arg	Val	Asp	Glu	Ile	Gln	His	Pro	Ile	Gly	Leu	Ala	Asn
			85						90					95	
Lys	Asn	Gly	Ala	Ser	Thr	Pro	Ser	Pro	Ser	Lys	Glu	Lys	Ser	Ile	Trp
		100						105					110		
Gly	Thr	Tyr	Ala	Arg	Gly	Ala	Val	Tyr	Ala	Leu	Gln	Ser	Ser	Lys	Lys
		115					120					125			
Asn	Leu	Lys	Gln	Gly	Ile	Ile	Gly	Tyr	Leu	Ser	Gly	Ser	Asn	Gly	Leu
	130					135					140				
Asp	Ser	Ser	Gly	Leu	Ser	Ser	Ser	Ala	Ala	Val	Gly	Val	Ala	Tyr	Leu
145				150						155					160
Leu	Ala	Leu	Glu	Asn	Ala	Asn	Glu	Leu	Thr	Val	Ser	Pro	Thr	Glu	Asn
			165						170					175	
Ile	Glu	Tyr	Asp	Arg	Leu	Ile	Glu	Asn	Val	Tyr	Leu	Gly	Leu	Arg	Asn
		180					185						190		
Gly	Ile	Leu	Asp	Gln	Ser	Ala	Ile	Leu	Leu	Ser	Asn	Tyr	Gly	Cys	Leu
		195					200					205			
Thr	Tyr	Met	Asp	Cys	Lys	Thr	Leu	Asp	His	Glu	Leu	Val	Gln	Ala	Pro
	210					215						220			
Glu	Leu	Glu	Lys	Pro	Phe	Arg	Ile	Leu	Leu	Ala	Phe	Ser	Gly	Leu	Arg
225				230						235					240
Gln	Ala	Leu	Thr	Thr	Asn	Pro	Gly	Tyr	Asn	Leu	Arg	Val	Ser	Glu	Cys
			245						250					255	
Gln	Glu	Ala	Ala	Lys	Val	Leu	Leu	Thr	Ala	Ser	Gly	Asn	Ser	Glu	Leu
		260						265					270		
Glu	Pro	Thr	Leu	Cys	Asn	Val	Glu	His	Ala	Val	Tyr	Glu	Ala	His	Lys
	275						280					285			
His	Glu	Leu	Lys	Pro	Val	Leu	Ala	Lys	Arg	Ala	Glu	His	Tyr	Phe	Ser
290					295						300				
Glu	Asn	Met	Arg	Val	Ile	Lys	Gly	Arg	Glu	Ala	Trp	Ala	Ser	Gly	Asn
305				310						315					320
Leu	Glu	Glu	Phe	Gly	Lys	Leu	Ile	Ser	Ala	Ser	Gly	Leu	Ser	Ser	Ile
			325						330					335	
Glu	Asn	Tyr	Glu	Cys	Gly	Ala	Glu	Pro	Leu	Ile	Gln	Leu	Tyr	Lys	Ile
		340						345					350		
Leu	Leu	Lys	Ala	Pro	Gly	Val	Tyr	Gly	Ala	Arg	Phe	Ser	Gly	Ala	Gly
		355					360					365			
Phe	Arg	Gly	Cys	Cys	Leu	Ala	Phe	Val	Asp	Ala	Val	Lys	Ala	Glu	Ala
370						375					380				
Ala	Ala	Ser	Tyr	Val	Lys	Asp	Glu	Tyr	Glu	Lys	Ala	Gln	Pro	Glu	Phe
385				390						395					400
Ala	Asn	Asn	Leu	Asn	Gly	Gly	Lys	Pro	Val	Leu	Ile	Cys	Glu	Ala	Gly
			405						410					415	
Asp	Ala	Ala	Arg	Val	Leu	Leu									
			420												

(2) INFORMATION FOR SEQ ID NO:70:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 405 amino acids
- (B) TYPE: amino acid

[illegible]

405

(2) INFORMATION FOR SEQ ID NO:71:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 371 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..371

(D) OTHER INFORMATION: / Ceres Seq. ID 1481547

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:71:

Met	Thr	Ile	Lys	Gly	Ile	Leu	Leu	Gly	Phe	Val	Pro	Ser	Gly	Asp	Thr
1			5					10						15	
Gln	Val	Gln	Leu	Arg	Ser	Ala	Gln	Phe	Glu	Gly	Glu	Val	Cys	Phe	Arg
			20					25					30		
Val	Asp	Glu	Ile	Gln	His	Pro	Ile	Gly	Leu	Ala	Asn	Lys	Asn	Gly	Ala
			35				40					45			
Ser	Thr	Pro	Ser	Pro	Ser	Lys	Glu	Lys	Ser	Ile	Trp	Gly	Thr	Tyr	Ala
			50			55					60				
Arg	Gly	Ala	Val	Tyr	Ala	Leu	Gln	Ser	Ser	Lys	Lys	Asn	Leu	Lys	Gln
65						70				75					80
Gly	Ile	Ile	Gly	Tyr	Leu	Ser	Gly	Ser	Asn	Gly	Leu	Asp	Ser	Ser	Gly
				85				90						95	
Leu	Ser	Ser	Ser	Ala	Ala	Val	Gly	Val	Ala	Tyr	Leu	Leu	Ala	Leu	Glu
				100				105					110		
Asn	Ala	Asn	Glu	Leu	Thr	Val	Ser	Pro	Thr	Glu	Asn	Ile	Glu	Tyr	Asp
				115			120					125			
Arg	Leu	Ile	Glu	Asn	Val	Tyr	Leu	Gly	Leu	Arg	Asn	Gly	Ile	Leu	Asp
				130			135				140				
Gln	Ser	Ala	Ile	Leu	Leu	Ser	Asn	Tyr	Gly	Cys	Leu	Thr	Tyr	Met	Asp
145					150					155					160
Cys	Lys	Thr	Leu	Asp	His	Glu	Leu	Val	Gln	Ala	Pro	Glu	Leu	Glu	Lys
				165					170					175	
Pro	Phe	Arg	Ile	Leu	Leu	Ala	Phe	Ser	Gly	Leu	Arg	Gln	Ala	Leu	Thr
				180				185					190		
Thr	Asn	Pro	Gly	Tyr	Asn	Leu	Arg	Val	Ser	Glu	Cys	Gln	Glu	Ala	Ala
				195			200					205			
Lys	Val	Leu	Leu	Thr	Ala	Ser	Gly	Asn	Ser	Glu	Leu	Glu	Pro	Thr	Leu
					215					220					
Cys	Asn	Val	Glu	His	Ala	Val	Tyr	Glu	Ala	His	Lys	His	Glu	Leu	Lys
225					230					235					240
Pro	Val	Leu	Ala	Lys	Arg	Ala	Glu	His	Tyr	Phe	Ser	Glu	Asn	Met	Arg
				245					250					255	
Val	Ile	Lys	Gly	Arg	Glu	Ala	Trp	Ala	Ser	Gly	Asn	Leu	Glu	Glu	Phe
				260				265					270		
Gly	Lys	Leu	Ile	Ser	Ala	Ser	Gly	Leu	Ser	Ser	Ile	Glu	Asn	Tyr	Glu
				275			280					285			
Cys	Gly	Ala	Glu	Pro	Leu	Ile	Gln	Leu	Tyr	Lys	Ile	Leu	Leu	Lys	Ala
					295					300					
Pro	Gly	Val	Tyr	Gly	Ala	Arg	Phe	Ser	Gly	Ala	Gly	Phe	Arg	Gly	Cys
305					310					315					320
Cys	Leu	Ala	Phe	Val	Asp	Ala	Val	Lys	Ala	Glu	Ala	Ala	Ala	Ser	Tyr
				325					330					335	
Val	Lys	Asp	Glu	Tyr	Glu	Lys	Ala	Gln	Pro	Glu	Phe	Ala	Asn	Asn	Leu
				340				345					350		
Asn	Gly	Gly	Lys	Pro	Val	Leu	Ile	Cys	Glu	Ala	Gly	Asp	Ala	Ala	Arg
				355			360					365			

Val Leu Leu
370

(2) INFORMATION FOR SEQ ID NO:72:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 915 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..915
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481564

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:72:

caaaagagag	aaaaggatgg	tcaataacgg	tccatgatct	ctccggttca	cccgtggcta	60
tggcctctat	ggtaacacct	ttcgtacat	ctcctgggtc	gaaccgggtg	actcgggtcaa	120
gtcccgagagc	gtggcttatt	cttcgacctg	acggttgac	atggaagcca	tggggaagac	180
tagaagcatg	gcgtgaggct	ggttactctg	acactctagg	ttatcgtttc	gagcttttcc	240
aagacgggtat	agccaccgca	gtttctgcat	cgctcgat	cagtttgaaa	aatggcgagg	300
gttttggtat	tgatgttacc	ggcgggtaca	gcacaacggc	gtctacgccg	acaacgagtc	360
ctcaaggaag	ctgggatctc	ggatccgggt	caagcgccg	ttcaagacc	gcgtcgagac	420
caggatcagg	atccgggtcg	gatttcggat	atctactacc	gcaacatccg	tctgcggccg	480
cgcaaacag	agggttcgtt	atgtcggcta	cggttgaagg	agttgggaaa	cgaagcaaac	540
cagaagtaga	agtcgggtgtg	acgcacgtga	catgtacgga	ggatgcagca	gcgcacgtgg	600
cattagctgc	ggcgggtgat	ctgagtttgg	atgcttgacg	gcttttctca	cacaagctaa	660
ggaaagagct	gagacagcaa	agccagcttg	gtgtcggttg	acttgtttcg	ctttgtcggt	720
ttaccaattc	atgagttgtc	ttccactcac	atttttttgg	tttgaatttt	ctattttttt	780
ctttttaaga	tagcgtagg	aattagccag	ccattttttt	gagaggtgga	tgatcatcatt	840
attaaaaatt	gttaatatct	ttctcagtac	agctaagaaa	tgacagtaac	aactaacaaa	900
caactcatta	tctcc					

(2) INFORMATION FOR SEQ ID NO:73:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 232 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..232
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481565

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:73:

Lys	Glu	Arg	Lys	Gly	Trp	Ser	Ile	Thr	Val	His	Asp	Leu	Ser	Gly	Ser
1			5					10						15	
Pro	Val	Ala	Met	Ala	Ser	Met	Val	Thr	Pro	Phe	Val	Pro	Ser	Pro	Gly
		20					25					30			
Ser	Asn	Arg	Val	Thr	Arg	Ser	Ser	Pro	Gly	Ala	Trp	Leu	Ile	Leu	Arg
	35					40					45				
Pro	Asp	Gly	Cys	Thr	Trp	Lys	Pro	Trp	Gly	Arg	Leu	Glu	Ala	Trp	Arg
	50					55				60					
Glu	Ala	Gly	Tyr	Ser	Asp	Thr	Leu	Gly	Tyr	Arg	Phe	Glu	Leu	Phe	Gln
	65				70				75					80	
Asp	Gly	Ile	Ala	Thr	Ala	Val	Ser	Ala	Ser	Ser	Ser	Ile	Ser	Leu	Lys
		85						90						95	
Asn	Gly	Gly	Ser	Phe	Val	Ile	Asp	Val	Thr	Gly	Gly	Thr	Ser	Thr	Thr
	100							105						110	
Ala	Ser	Thr	Pro	Thr	Thr	Ser	Pro	Gln	Gly	Ser	Trp	Asp	Leu	Gly	Ser
	115					120						125			
Gly	Ser	Ser	Ala	Gly	Ser	Arg	Pro	Ala	Ser	Arg	Pro	Gly	Ser	Gly	Ser

130		135		140											
Gly	Ser	Asp	Phe	Gly	Tyr	Leu	Leu	Pro	Gln	His	Pro	Ser	Ala	Ala	Ala
145					150					155					160
Gln	Asn	Arg	Gly	Phe	Val	Met	Ser	Ala	Thr	Val	Glu	Gly	Val	Gly	Lys
				165					170						175
Arg	Ser	Lys	Pro	Glu	Val	Glu	Val	Gly	Val	Thr	His	Val	Thr	Cys	Thr
			180					185						190	
Glu	Asp	Ala	Ala	Ala	His	Val	Ala	Leu	Ala	Ala	Ala	Val	Asp	Leu	Ser
		195					200					205			
Leu	Asp	Ala	Cys	Arg	Leu	Phe	Ser	His	Lys	Leu	Arg	Lys	Glu	Leu	Arg
	210					215					220				
Gln	Gln	Ser	Gln	Leu	Gly	Val	Val								
225					230										

(2) INFORMATION FOR SEQ ID NO:74:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 213 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..213
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481566

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:74:

Met	Ala	Ser	Met	Val	Thr	Pro	Phe	Val	Pro	Ser	Pro	Gly	Ser	Asn	Arg
1			5					10						15	
Val	Thr	Arg	Ser	Ser	Pro	Gly	Ala	Trp	Leu	Ile	Leu	Arg	Pro	Asp	Gly
			20					25					30		
Cys	Thr	Trp	Lys	Pro	Trp	Gly	Arg	Leu	Glu	Ala	Trp	Arg	Glu	Ala	Gly
		35				40					45				
Tyr	Ser	Asp	Thr	Leu	Gly	Tyr	Arg	Phe	Glu	Leu	Phe	Gln	Asp	Gly	Ile
	50				55						60				
Ala	Thr	Ala	Val	Ser	Ala	Ser	Ser	Ser	Ile	Ser	Leu	Lys	Asn	Gly	Gly
	65				70					75				80	
Ser	Phe	Val	Ile	Asp	Val	Thr	Gly	Gly	Thr	Ser	Thr	Thr	Ala	Ser	Thr
			85					90					95		
Pro	Thr	Thr	Ser	Pro	Gln	Gly	Ser	Trp	Asp	Leu	Gly	Ser	Gly	Ser	Ser
		100					105						110		
Ala	Gly	Ser	Arg	Pro	Ala	Ser	Arg	Pro	Gly	Ser	Gly	Ser	Gly	Ser	Asp
		115					120					125			
Phe	Gly	Tyr	Leu	Leu	Pro	Gln	His	Pro	Ser	Ala	Ala	Ala	Gln	Asn	Arg
	130					135					140				
Gly	Phe	Val	Met	Ser	Ala	Thr	Val	Glu	Gly	Val	Gly	Lys	Arg	Ser	Lys
	145				150					155				160	
Pro	Glu	Val	Glu	Val	Gly	Val	Thr	His	Val	Thr	Cys	Thr	Glu	Asp	Ala
			165					170						175	
Ala	Ala	His	Val	Ala	Leu	Ala	Ala	Ala	Val	Asp	Leu	Ser	Leu	Asp	Ala
		180						185					190		
Cys	Arg	Leu	Phe	Ser	His	Lys	Leu	Arg	Lys	Glu	Leu	Arg	Gln	Gln	Ser
		195					200						205		
Gln	Leu	Gly	Val	Val											
	210														

(2) INFORMATION FOR SEQ ID NO:75:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 210 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..210
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481567
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:75:

```
Met Val Thr Pro Phe Val Pro Ser Pro Gly Ser Asn Arg Val Thr Arg
1          5          10          15
Ser Ser Pro Gly Ala Trp Leu Ile Leu Arg Pro Asp Gly Cys Thr Trp
          20          25          30
Lys Pro Trp Gly Arg Leu Glu Ala Trp Arg Glu Ala Gly Tyr Ser Asp
          35          40          45
Thr Leu Gly Tyr Arg Phe Glu Leu Phe Gln Asp Gly Ile Ala Thr Ala
          50          55          60
Val Ser Ala Ser Ser Ser Ile Ser Leu Lys Asn Gly Gly Ser Phe Val
65          70          75          80
Ile Asp Val Thr Gly Gly Thr Ser Thr Thr Ala Ser Thr Pro Thr Thr
          85          90          95
Ser Pro Gln Gly Ser Trp Asp Leu Gly Ser Gly Ser Ser Ala Gly Ser
          100         105         110
Arg Pro Ala Ser Arg Pro Gly Ser Gly Ser Gly Ser Asp Phe Gly Tyr
          115         120         125
Leu Leu Pro Gln His Pro Ser Ala Ala Ala Gln Asn Arg Gly Phe Val
130         135         140
Met Ser Ala Thr Val Glu Gly Val Gly Lys Arg Ser Lys Pro Glu Val
145         150         155         160
Glu Val Gly Val Thr His Val Thr Cys Thr Glu Asp Ala Ala Ala His
          165         170         175
Val Ala Leu Ala Ala Ala Val Asp Leu Ser Leu Asp Ala Cys Arg Leu
          180         185         190
Phe Ser His Lys Leu Arg Lys Glu Leu Arg Gln Gln Ser Gln Leu Gly
195         200         205
Val Val
210
```

- (2) INFORMATION FOR SEQ ID NO:76:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 1330 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: DNA (genomic)
 (ix) FEATURE:

- (A) NAME/KEY: -
 (B) LOCATION: 1..1330
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481580

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:76:

```
atcccaaagt ctcctaaggc gagaagggat ccggtcgctg cactaggttt ccttttcttt 60
tttttttttg gtttcaaatt ttttcatcat atctcctgaa aatcttcttc attcgctcc 120
aattttgctc atgttcgtcc gatccaacat tcttcgtgct ctgattttta ctgtgctgga 180
aacagttacc tcctcttgat tcagttttga ttcttcaaag cctcagagat aatttggttt 240
tctctaattc ctgtgaagga gaaaacttgc ttggagatca aaatgatgca ttcaagctgc 300
aaaggacttg tgctgcttct attcttatth gtcgttgat tcattggaaa caccgatgcg 360
aatgctcagt gggagggttc acataaagta agagcttctc cccatgaaaa catgggacgt 420
aatgttattg acggaagtgg ttagagaaaa acgttacatg acattggaat gggtgaaaag 480
agaggcactc acaacaaagt ttcagtctca acagttgcgt tgttcacctt ggctatggct 540
gctgccactg ggtaggtgct tgtgcccttc ttctttgttg agcttgatcc tcaatgggct 600
ggaatttgca atggcatggc tgctggttga tgttggccgc tagctttgat cttgttaagg 660
aagggcagga acatggctct ggaaactggg ttgttactgg gatcctagcc ggtgctttgt 720
tcatttggtc ctgtaagcag attcttgaac aatatggtga agttagtatg ctggatatta 780
```

```
aaggcgcaga tgcaactaaa gttgttctcg tcataggaat tatgacactt cattctttcg      840
gggaaggatc aggggttggg gtatcattcg ctggctcaaa aggttttagc caagggcttc      900
tggtcacttt ggccatagct gttcataaca ttccagaagg gttggctgtt agcatggtgt      960
tggcatcaag ggggtgtctt ccacaaaatg ccatgctctg gagtataata acatccttac     1020
ctcagcctct cgtcgccgtg ccagcttttt tatgcgctga tgcgttcagc aagtttttgc     1080
ctttttgcac tggatttgct gccggatgca tgatttggat ggttattgct gaagtgcctc     1140
ctgatgcttt taagggaagcg tctccttcgc aagtggcatc tgcagccacc atatcagtag     1200
catccatgga agctcttagc actcttttcg agagtttcac acatgattac aactcagagg     1260
atgcttctgg cttcttcggt tcaactcctt ttgggtctggg tccattgctt gggggagtat     1320
ttctgggtgc
```

(2) INFORMATION FOR SEQ ID NO:77:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 245 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..245
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481581

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:77:

```
Met Gly Trp Asn Leu Gln Trp His Gly Cys Trp Leu Met Leu Ala Ala
1          5          10          15
Ser Phe Asp Leu Val Lys Glu Gly Gln Glu His Gly Ser Gly Asn Trp
20          25          30
Val Val Thr Gly Ile Leu Ala Gly Ala Leu Phe Ile Trp Leu Cys Lys
35          40          45
Gln Ile Leu Glu Gln Tyr Gly Glu Val Ser Met Leu Asp Ile Lys Gly
50          55          60
Ala Asp Ala Thr Lys Val Val Leu Val Ile Gly Ile Met Thr Leu His
65          70          75          80
Ser Phe Gly Glu Gly Ser Gly Val Gly Val Ser Phe Ala Gly Ser Lys
85          90          95
Gly Phe Ser Gln Gly Leu Leu Val Thr Leu Ala Ile Ala Val His Asn
100          105          110
Ile Pro Glu Gly Leu Ala Val Ser Met Val Leu Ala Ser Arg Gly Val
115          120          125
Ser Pro Gln Asn Ala Met Leu Trp Ser Ile Ile Thr Ser Leu Pro Gln
130          135          140
Pro Leu Val Ala Val Pro Ala Phe Leu Cys Ala Asp Ala Phe Ser Lys
145          150          155          160
Phe Leu Pro Phe Cys Thr Gly Phe Ala Ala Gly Cys Met Ile Trp Met
165          170          175
Val Ile Ala Glu Val Leu Pro Asp Ala Phe Lys Glu Ala Ser Pro Ser
180          185          190
Gln Val Ala Ser Ala Ala Thr Ile Ser Val Ala Ser Met Glu Ala Leu
195          200          205
Ser Thr Leu Phe Glu Ser Phe Thr His Asp Tyr Asn Ser Glu Asp Ala
210          215          220
Ser Gly Phe Phe Val Ser Leu Leu Phe Gly Leu Gly Pro Leu Leu Gly
225          230          235          240
Gly Val Phe Leu Val
245
```

(2) INFORMATION FOR SEQ ID NO:78:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 233 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:

(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
(A) NAME/KEY: peptide
(B) LOCATION: 1..233
(D) OTHER INFORMATION: / Ceres Seq. ID 1481582

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:78:

Met	Leu	Ala	Ala	Ser	Phe	Asp	Leu	Val	Lys	Glu	Gly	Gln	Glu	His	Gly	
1				5					10					15		
Ser	Gly	Asn	Trp	Val	Val	Thr	Gly	Ile	Leu	Ala	Gly	Ala	Leu	Phe	Ile	
			20					25					30			
Trp	Leu	Cys	Lys	Gln	Ile	Leu	Glu	Gln	Tyr	Gly	Glu	Val	Ser	Met	Leu	
			35				40					45				
Asp	Ile	Lys	Gly	Ala	Asp	Ala	Thr	Lys	Val	Val	Leu	Val	Ile	Gly	Ile	
	50					55					60					
Met	Thr	Leu	His	Ser	Phe	Gly	Glu	Gly	Ser	Gly	Val	Gly	Val	Ser	Phe	
65					70				75						80	
Ala	Gly	Ser	Lys	Gly	Phe	Ser	Gln	Gly	Leu	Leu	Val	Thr	Leu	Ala	Ile	
				85					90					95		
Ala	Val	His	Asn	Ile	Pro	Glu	Gly	Leu	Ala	Val	Ser	Met	Val	Leu	Ala	
			100					105					110			
Ser	Arg	Gly	Val	Ser	Pro	Gln	Asn	Ala	Met	Leu	Trp	Ser	Ile	Ile	Thr	
	115						120						125			
Ser	Leu	Pro	Gln	Pro	Leu	Val	Ala	Val	Pro	Ala	Phe	Leu	Cys	Ala	Asp	
	130					135					140					
Ala	Phe	Ser	Lys	Phe	Leu	Pro	Phe	Cys	Thr	Gly	Phe	Ala	Ala	Gly	Cys	
145					150					155					160	
Met	Ile	Trp	Met	Val	Ile	Ala	Glu	Val	Leu	Pro	Asp	Ala	Phe	Lys	Glu	
				165					170					175		
Ala	Ser	Pro	Ser	Gln	Val	Ala	Ser	Ala	Ala	Thr	Ile	Ser	Val	Ala	Ser	
			180					185					190			
Met	Glu	Ala	Leu	Ser	Thr	Leu	Phe	Glu	Ser	Phe	Thr	His	Asp	Tyr	Asn	
	195						200					205				
Ser	Glu	Asp	Ala	Ser	Gly	Phe	Phe	Val	Ser	Leu	Leu	Phe	Gly	Leu	Gly	
	210					215					220					
Pro	Leu	Leu	Gly	Gly	Val	Phe	Leu	Val								
225					230											

(2) INFORMATION FOR SEQ ID NO:79:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 187 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
(A) NAME/KEY: peptide
(B) LOCATION: 1..187
(D) OTHER INFORMATION: / Ceres Seq. ID 1481583

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:79:

Met	Leu	Asp	Ile	Lys	Gly	Ala	Asp	Ala	Thr	Lys	Val	Val	Leu	Val	Ile	
1				5					10					15		
Gly	Ile	Met	Thr	Leu	His	Ser	Phe	Gly	Glu	Gly	Ser	Gly	Val	Gly	Val	
			20						25				30			
Ser	Phe	Ala	Gly	Ser	Lys	Gly	Phe	Ser	Gln	Gly	Leu	Leu	Val	Thr	Leu	
		35					40					45				
Ala	Ile	Ala	Val	His	Asn	Ile	Pro	Glu	Gly	Leu	Ala	Val	Ser	Met	Val	
	50					55					60					
Leu	Ala	Ser	Arg	Gly	Val	Ser	Pro	Gln	Asn	Ala	Met	Leu	Trp	Ser	Ile	
65					70					75					80	

Ile Thr Ser Leu Pro Gln Pro Leu Val Ala Val Pro Ala Phe Leu Cys
85 90 95
Ala Asp Ala Phe Ser Lys Phe Leu Pro Phe Cys Thr Gly Phe Ala Ala
100 105 110
Gly Cys Met Ile Trp Met Val Ile Ala Glu Val Leu Pro Asp Ala Phe
115 120 125
Lys Glu Ala Ser Pro Ser Gln Val Ala Ser Ala Ala Thr Ile Ser Val
130 135 140
Ala Ser Met Glu Ala Leu Ser Thr Leu Phe Glu Ser Phe Thr His Asp
145 150 155 160
Tyr Asn Ser Glu Asp Ala Ser Gly Phe Phe Val Ser Leu Leu Phe Gly
165 170 175
Leu Gly Pro Leu Leu Gly Gly Val Phe Leu Val
180 185

(2) INFORMATION FOR SEQ ID NO:80:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1180 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1180
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481596

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:80:

acaattctaa aaccctaatac tcacaaaaaa ccctaattctc acaaaaaccc tcctctcttc	60
atcgacatct ctcttttcaact gcttcaatgg cgtcttttga gcgttttgac gacatgtgtg	120
acctgagatt gaaacctaac attctccgaa accttctctc cgaatatgtt cccaacgaga	180
agcagcctct caccaacttt ctatcactct ccaaggttgt atcaaccatc tccacacaca	240
agctcttatac tgagtctcct ccagcttcaa ttgaccagaa gcttcatgct aaatcgaaat	300
cagccgttga tgattgggtt gctagattat cagctttgat ttcttctgat atgccggata	360
aaagctgggt gggatattgt ttgattggag taacatgtca agaattgtagc tcagatcggt	420
tctttaagtc atactctggt tggtttaaca gtttattatc acatcttaag aatccagcaa	480
gttctagaat tgtccgagtg gcttcatgta cctcaatctc tgatctactt acaaggctgt	540
ctagattttc gaatacgaag aaagatgcag ttccacacgc ttcgaaacta atcctgccta	600
tcattaaatt attggatgaa gattcttcag aagcactatt ggaaggcatt gtccatctgc	660
taagtacaat tgtactcttg tttcctgctg ccttccacag taattatgac aagattgaag	720
ccgctattgc ctccaaaata ttttcggcga aaaccagttc taatatgtta aagaaatttg	780
cccactttct agcattgctc cccaaagcta aaggtgacga gggcacctgg tccttgatga	840
tgcaaaagct gctgatattc ataaacgtac atttaaataa ttttttccaa ggtctagaag	900
aagaaacaaa aggaacaaaa gcaatccaac gattgactcc tcctggaaaa gactctcctt	960
tgccctcctg aggtcaaaat gggggattgg atgatgcac atggaactct gaacaattga	1020
ttgtatccag agtttctgca cttatgttct gcacctcaac gatgttaact acctcgtaca	1080
aatccaagat taatattcca gttggctcat tgttatccct tgttgagcga gtgctgttg	1140
tgaacggctc tctacctcga gccatgtcac ccttcatgac	

(2) INFORMATION FOR SEQ ID NO:81:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 392 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..392
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481597

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:81:

Asn Ser Lys Thr Leu Ile Ser Gln Lys Thr Leu Ile Ser Gln Lys Pro

1	5	10	15
Ser Ser Leu His Arg His Leu Ser Phe Thr Ala Ser Met Ala Ser Phe			
	20	25	30
Glu Arg Phe Asp Asp Met Cys Asp Leu Arg Leu Lys Pro Asn Ile Leu			
	35	40	45
Arg Asn Leu Leu Ser Glu Tyr Val Pro Asn Glu Lys Gln Pro Leu Thr			
	50	55	60
Asn Phe Leu Ser Leu Ser Lys Val Val Ser Thr Ile Ser Thr His Lys			
	65	70	75
Leu Leu Ser Glu Ser Pro Pro Ala Ser Ile Asp Gln Lys Leu His Ala			
	85	90	95
Lys Ser Lys Ser Ala Val Asp Asp Trp Val Ala Arg Leu Ser Ala Leu			
	100	105	110
Ile Ser Ser Asp Met Pro Asp Lys Ser Trp Val Gly Ile Cys Leu Ile			
	115	120	125
Gly Val Thr Cys Gln Glu Cys Ser Ser Asp Arg Phe Phe Lys Ser Tyr			
	130	135	140
Ser Val Trp Phe Asn Ser Leu Leu Ser His Leu Lys Asn Pro Ala Ser			
	145	150	155
Ser Arg Ile Val Arg Val Ala Ser Cys Thr Ser Ile Ser Asp Leu Leu			
	165	170	175
Thr Arg Leu Ser Arg Phe Ser Asn Thr Lys Lys Asp Ala Val Ser His			
	180	185	190
Ala Ser Lys Leu Ile Leu Pro Ile Lys Leu Leu Asp Glu Asp Ser			
	195	200	205
Ser Glu Ala Leu Leu Glu Gly Ile Val His Leu Leu Ser Thr Ile Val			
	210	215	220
Leu Leu Phe Pro Ala Ala Phe His Ser Asn Tyr Asp Lys Ile Glu Ala			
	225	230	235
Ala Ile Ala Ser Lys Ile Phe Ser Ala Lys Thr Ser Ser Asn Met Leu			
	245	250	255
Lys Lys Phe Ala His Phe Leu Ala Leu Leu Pro Lys Ala Lys Gly Asp			
	260	265	270
Glu Gly Thr Trp Ser Leu Met Met Gln Lys Leu Leu Ile Ser Ile Asn			
	275	280	285
Val His Leu Asn Asn Phe Phe Gln Gly Leu Glu Glu Glu Thr Lys Gly			
	290	295	300
Thr Lys Ala Ile Gln Arg Leu Thr Pro Pro Gly Lys Asp Ser Pro Leu			
	305	310	315
Pro Leu Gly Gly Gln Asn Gly Gly Leu Asp Asp Ala Ser Trp Asn Ser			
	325	330	335
Glu Gln Leu Ile Val Ser Arg Val Ser Ala Leu Met Phe Cys Thr Ser			
	340	345	350
Thr Met Leu Thr Thr Ser Tyr Lys Ser Lys Ile Asn Ile Pro Val Gly			
	355	360	365
Ser Leu Leu Ser Leu Val Glu Arg Val Leu Leu Val Asn Gly Ser Leu			
	370	375	380
Pro Arg Ala Met Ser Pro Phe Met			
	385	390	

(2) INFORMATION FOR SEQ ID NO:82:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 364 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..364

(D) OTHER INFORMATION: / Ceres Seq. ID 1481598

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:82:

Met	Ala	Ser	Phe	Glu	Arg	Phe	Asp	Asp	Met	Cys	Asp	Leu	Arg	Leu	Lys
1				5				10						15	
Pro	Asn	Ile	Leu	Arg	Asn	Leu	Leu	Ser	Glu	Tyr	Val	Pro	Asn	Glu	Lys
			20					25					30		
Gln	Pro	Leu	Thr	Asn	Phe	Leu	Ser	Leu	Ser	Lys	Val	Val	Ser	Thr	Ile
		35					40					45			
Ser	Thr	His	Lys	Leu	Leu	Ser	Glu	Ser	Pro	Pro	Ala	Ser	Ile	Asp	Gln
	50					55				60					
Lys	Leu	His	Ala	Lys	Ser	Lys	Ser	Ala	Val	Asp	Asp	Trp	Val	Ala	Arg
65					70					75					80
Leu	Ser	Ala	Leu	Ile	Ser	Ser	Asp	Met	Pro	Asp	Lys	Ser	Trp	Val	Gly
			85					90					95		
Ile	Cys	Leu	Ile	Gly	Val	Thr	Cys	Gln	Glu	Cys	Ser	Ser	Asp	Arg	Phe
		100						105					110		
Phe	Lys	Ser	Tyr	Ser	Val	Trp	Phe	Asn	Ser	Leu	Leu	Ser	His	Leu	Lys
		115					120					125			
Asn	Pro	Ala	Ser	Ser	Arg	Ile	Val	Arg	Val	Ala	Ser	Cys	Thr	Ser	Ile
	130					135						140			
Ser	Asp	Leu	Leu	Thr	Arg	Leu	Ser	Arg	Phe	Ser	Asn	Thr	Lys	Lys	Asp
145					150					155					160
Ala	Val	Ser	His	Ala	Ser	Lys	Leu	Ile	Leu	Pro	Ile	Ile	Lys	Leu	Leu
			165						170					175	
Asp	Glu	Asp	Ser	Ser	Glu	Ala	Leu	Leu	Glu	Gly	Ile	Val	His	Leu	Leu
		180						185					190		
Ser	Thr	Ile	Val	Leu	Leu	Phe	Pro	Ala	Ala	Phe	His	Ser	Asn	Tyr	Asp
	195						200					205			
Lys	Ile	Glu	Ala	Ala	Ile	Ala	Ser	Lys	Ile	Phe	Ser	Ala	Lys	Thr	Ser
	210					215					220				
Ser	Asn	Met	Leu	Lys	Lys	Phe	Ala	His	Phe	Leu	Ala	Leu	Leu	Pro	Lys
225					230					235					240
Ala	Lys	Gly	Asp	Glu	Gly	Thr	Trp	Ser	Leu	Met	Met	Gln	Lys	Leu	Leu
			245						250					255	
Ile	Ser	Ile	Asn	Val	His	Leu	Asn	Asn	Phe	Phe	Gln	Gly	Leu	Glu	Glu
		260						265					270		
Glu	Thr	Lys	Gly	Thr	Lys	Ala	Ile	Gln	Arg	Leu	Thr	Pro	Pro	Gly	Lys
	275						280					285			
Asp	Ser	Pro	Leu	Pro	Leu	Gly	Gly	Gln	Asn	Gly	Gly	Leu	Asp	Asp	Ala
	290					295					300				
Ser	Trp	Asn	Ser	Glu	Gln	Leu	Ile	Val	Ser	Arg	Val	Ser	Ala	Leu	Met
305					310					315					320
Phe	Cys	Thr	Ser	Thr	Met	Leu	Thr	Thr	Ser	Tyr	Lys	Ser	Lys	Ile	Asn
			325						330					335	
Ile	Pro	Val	Gly	Ser	Leu	Leu	Ser	Leu	Val	Glu	Arg	Val	Leu	Leu	Val
		340						345					350		
Asn	Gly	Ser	Leu	Pro	Arg	Ala	Met	Ser	Pro	Phe	Met				
		355					360								

(2) INFORMATION FOR SEQ ID NO:83:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 355 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..355

(D) OTHER INFORMATION: / Ceres Seq. ID 1481599

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:83:

Met	Cys	Asp	Leu	Arg	Leu	Lys	Pro	Asn	Ile	Leu	Arg	Asn	Leu	Leu	Ser
1			5					10					15		
Glu	Tyr	Val	Pro	Asn	Glu	Lys	Gln	Pro	Leu	Thr	Asn	Phe	Leu	Ser	Leu
			20					25					30		
Ser	Lys	Val	Val	Ser	Thr	Ile	Ser	Thr	His	Lys	Leu	Leu	Ser	Glu	Ser
		35					40					45			
Pro	Pro	Ala	Ser	Ile	Asp	Gln	Lys	Leu	His	Ala	Lys	Ser	Lys	Ser	Ala
		50				55					60				
Val	Asp	Asp	Trp	Val	Ala	Arg	Leu	Ser	Ala	Leu	Ile	Ser	Ser	Asp	Met
65					70				75					80	
Pro	Asp	Lys	Ser	Trp	Val	Gly	Ile	Cys	Leu	Ile	Gly	Val	Thr	Cys	Gln
			85					90						95	
Glu	Cys	Ser	Ser	Asp	Arg	Phe	Phe	Lys	Ser	Tyr	Ser	Val	Trp	Phe	Asn
			100					105					110		
Ser	Leu	Leu	Ser	His	Leu	Lys	Asn	Pro	Ala	Ser	Ser	Arg	Ile	Val	Arg
		115					120					125			
Val	Ala	Ser	Cys	Thr	Ser	Ile	Ser	Asp	Leu	Leu	Thr	Arg	Leu	Ser	Arg
		130				135					140				
Phe	Ser	Asn	Thr	Lys	Lys	Asp	Ala	Val	Ser	His	Ala	Ser	Lys	Leu	Ile
145					150					155				160	
Leu	Pro	Ile	Ile	Lys	Leu	Leu	Asp	Glu	Asp	Ser	Ser	Glu	Ala	Leu	Leu
			165					170						175	
Glu	Gly	Ile	Val	His	Leu	Leu	Ser	Thr	Ile	Val	Leu	Leu	Phe	Pro	Ala
			180					185					190		
Ala	Phe	His	Ser	Asn	Tyr	Asp	Lys	Ile	Glu	Ala	Ala	Ile	Ala	Ser	Lys
		195					200					205			
Ile	Phe	Ser	Ala	Lys	Thr	Ser	Ser	Asn	Met	Leu	Lys	Lys	Phe	Ala	His
		210				215					220				
Phe	Leu	Ala	Leu	Leu	Pro	Lys	Ala	Lys	Gly	Asp	Glu	Gly	Thr	Trp	Ser
225					230				235					240	
Leu	Met	Met	Gln	Lys	Leu	Leu	Ile	Ser	Ile	Asn	Val	His	Leu	Asn	Asn
			245					250					255		
Phe	Phe	Gln	Gly	Leu	Glu	Glu	Glu	Thr	Lys	Gly	Thr	Lys	Ala	Ile	Gln
		260						265					270		
Arg	Leu	Thr	Pro	Pro	Gly	Lys	Asp	Ser	Pro	Leu	Pro	Leu	Gly	Gly	Gln
		275					280					285			
Asn	Gly	Gly	Leu	Asp	Asp	Ala	Ser	Trp	Asn	Ser	Glu	Gln	Leu	Ile	Val
		290				295					300				
Ser	Arg	Val	Ser	Ala	Leu	Met	Phe	Cys	Thr	Ser	Thr	Met	Leu	Thr	Thr
305					310					315				320	
Ser	Tyr	Lys	Ser	Lys	Ile	Asn	Ile	Pro	Val	Gly	Ser	Leu	Leu	Ser	Leu
			325					330					335		
Val	Glu	Arg	Val	Leu	Leu	Val	Asn	Gly	Ser	Leu	Pro	Arg	Ala	Met	Ser
			340				345					350			
Pro	Phe	Met													
		355													

(2) INFORMATION FOR SEQ ID NO:84:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1724 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1724
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481613

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:84:


```
attcaacctc tctacttcag tttctctgtc cccatttttc atctgagagt taaaactgta 60
acctcaaaat ctgagataaa gtcaaaaaaa aaaccccagt ttatgattct cattttctct 120
ttataatcga aagcttcgat ttttaacaaa acccagaatc tgtgttcttg tttttttttt 180
tttttggtga gagttatctt tttttttttt ggaatattgg gtgagaatct gagtaatggg 240
atttacataa aatattctat gtaaabacta aaataatctg gaaattatta aatttccaaa 300
ctttgtgttc cattttgtgg actcaaattt gtttataaag atctcaaadc agagagattg 360
agacgaccaa gaacaagcag aagaagaaga agaattgaga gaatgtgggt gtggtcttct 420
tcaactaaag gtcgttcgaa tctggagagg tttcttttag gaatcactcc taagcctcct 480
tccttctctc ttcttcagga acagggaaaag gaggagattg agtatttcag gcttgatgat 540
ctctgggatt gttatgatga gatgagtgcg tatggctttg gcacacaggt tgatttaaac 600
aatggcgaaa ccgttatgca gtactacgtc ccgtacctat ccgctatcca aatccacact 660
aacaaccccg ccttgctttc caggaaccag aatgagggtg ctgaatctga gagtagcgag 720
ggttgaggcg atagtggagag tgaaggattg ttgtcaaggc caatgagcaa tgattcaagc 780
aaaacatggg atgctgtctc tgaagattcg gttttcgatc cggatgggtc accgttgctg 840
aaagatagac ttggtaacct tgactttaag tacattgaaa gagatcctcc gcacaagcgg 900
attcccttaa ccgacaagat aaacgtattg gtggagaaat atccgggact catgacctta 960
aggagtgtcg acatgtctcc tgcaagttgg atggctgttg cgtggtaccc gatataccac 1020
atcccaacct gcaggaacga gaaagatttg acgacaggct tcctaactta tcatactcta 1080
tcttcgtctt ttcaagataa tgtgggtggaa ggagatcaaa gcaacaacaa tgaagaaaca 1140
gagttttgtg aagattccgt aataaacaag agaatgccat tgcctccggt tgggtgaaca 1200
acttacaaaa tgcaaggaga tctttggggg aagacggggt ttgaccagga ccggttgctt 1260
tatcttcaaa gcgctgcgga ttcattggctg aaacagctca atgttgatca ccatgactat 1320
aacttcttcc ttaactcgag cttctaaaga tcaatcgggt cgttcgtatg tttatccttc 1380
tccaaacctt aaacaaaaaa aaaaagacct cataaccctt tttctttgtt gttttcaagc 1440
tccttttggt tctctgtggt ttttgttctt tttgtttttg tctggctcgt tgtgtgtttt 1500
taggtagcaa ccgccatcgc ggagtttttt ctccttttgc aagccaatca tggaagtttc 1560
taagaagaaa acagagcttt tttttctttt tttttaacgg tgttgagaaa acaagaaagt 1620
tgttttcttt tcttggttga gagatcatgt aaattgacct tgaacagagg actctgtttt 1680
gtacttttct gtctaaaata tataaaaaaa tctgtctttc ttgt
```

(2) INFORMATION FOR SEQ ID NO:85:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 314 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..314
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481614

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:85:

```
Met Trp Trp Trp Ser Ser Thr Lys Gly Arg Ser Asn Leu Glu Arg
1           5           10           15
Phe Leu Leu Gly Ile Thr Pro Lys Pro Pro Ser Phe Ser Leu Pro Gln
20           25           30
Glu Gln Gly Lys Glu Glu Ile Glu Tyr Phe Arg Leu Asp Asp Leu Trp
35           40           45
Asp Cys Tyr Asp Glu Met Ser Ala Tyr Gly Phe Gly Thr Gln Val Asp
50           55           60
Leu Asn Asn Gly Glu Thr Val Met Gln Tyr Tyr Val Pro Tyr Leu Ser
65           70           75           80
Ala Ile Gln Ile His Thr Asn Lys Pro Ala Leu Leu Ser Arg Asn Gln
85           90           95
Asn Glu Val Ala Glu Ser Glu Ser Ser Glu Gly Trp Ser Asp Ser Glu
100          105          110
Ser Glu Lys Leu Leu Ser Arg Ser Met Ser Asn Asp Ser Ser Lys Thr
115          120          125
Trp Asp Ala Val Ser Glu Asp Ser Val Phe Asp Pro Asp Gly Ser Pro
130          135          140
```

Leu Leu Lys Asp Arg Leu Gly Asn Leu Asp Phe Lys Tyr Ile Glu Arg
145 150 155 160
Asp Pro Pro His Lys Arg Ile Pro Leu Thr Asp Lys Ile Asn Val Leu
165 170 175
Val Glu Lys Tyr Pro Gly Leu Met Thr Leu Arg Ser Val Asp Met Ser
180 185 190
Pro Ala Ser Trp Met Ala Val Ala Trp Tyr Pro Ile Tyr His Ile Pro
195 200 205
Thr Cys Arg Asn Glu Lys Asp Leu Thr Thr Gly Phe Leu Thr Tyr His
210 215 220
Thr Leu Ser Ser Ser Phe Gln Asp Asn Val Val Glu Gly Asp Gln Ser
225 230 235 240
Asn Asn Asn Glu Glu Thr Glu Phe Cys Glu Asp Ser Val Ile Asn Lys
245 250 255
Arg Met Pro Leu Pro Pro Phe Gly Val Thr Thr Tyr Lys Met Gln Gly
260 265 270
Asp Leu Trp Gly Lys Thr Gly Phe Asp Gln Asp Arg Leu Leu Tyr Leu
275 280 285
Gln Ser Ala Ala Asp Ser Trp Leu Lys Gln Leu Asn Val Asp His His
290 295 300
Asp Tyr Asn Phe Phe Leu Asn Ser Ser Phe
305 310

(2) INFORMATION FOR SEQ ID NO:86:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 261 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..261

(D) OTHER INFORMATION: / Ceres Seq. ID 1481615

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:86:

Met Ser Ala Tyr Gly Phe Gly Thr Gln Val Asp Leu Asn Asn Gly Glu
1 5 10 15
Thr Val Met Gln Tyr Tyr Val Pro Tyr Leu Ser Ala Ile Gln Ile His
20 25 30
Thr Asn Lys Pro Ala Leu Leu Ser Arg Asn Gln Asn Glu Val Ala Glu
35 40 45
Ser Glu Ser Ser Glu Gly Trp Ser Asp Ser Glu Ser Glu Lys Leu Leu
50 55 60
Ser Arg Ser Met Ser Asn Asp Ser Ser Lys Thr Trp Asp Ala Val Ser
65 70 75 80
Glu Asp Ser Val Phe Asp Pro Asp Gly Ser Pro Leu Leu Lys Asp Arg
85 90 95
Leu Gly Asn Leu Asp Phe Lys Tyr Ile Glu Arg Asp Pro Pro His Lys
100 105 110
Arg Ile Pro Leu Thr Asp Lys Ile Asn Val Leu Val Glu Lys Tyr Pro
115 120 125
Gly Leu Met Thr Leu Arg Ser Val Asp Met Ser Pro Ala Ser Trp Met
130 135 140
Ala Val Ala Trp Tyr Pro Ile Tyr His Ile Pro Thr Cys Arg Asn Glu
145 150 155 160
Lys Asp Leu Thr Thr Gly Phe Leu Thr Tyr His Thr Leu Ser Ser Ser
165 170 175
Phe Gln Asp Asn Val Val Glu Gly Asp Gln Ser Asn Asn Asn Glu Glu
180 185 190
Thr Glu Phe Cys Glu Asp Ser Val Ile Asn Lys Arg Met Pro Leu Pro

195	200	205
Pro Phe Gly Val Thr Thr Tyr Lys Met Gln Gly Asp Leu Trp Gly Lys		
210	215	220
Thr Gly Phe Asp Gln Asp Arg Leu Leu Tyr Leu Gln Ser Ala Ala Asp		
225	230	235
Ser Trp Leu Lys Gln Leu Asn Val Asp His His Asp Tyr Asn Phe Phe		
	245	250
Leu Asn Ser Ser Phe		255
260		

(2) INFORMATION FOR SEQ ID NO:87:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 243 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..243

(D) OTHER INFORMATION: / Ceres Seq. ID 1481616

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:87:

Met Gln Tyr Tyr Val Pro Tyr Leu Ser Ala Ile Gln Ile His Thr Asn		
1	5	10
Lys Pro Ala Leu Ser Arg Asn Gln Asn Glu Val Ala Glu Ser Glu		
	20	25
Ser Ser Glu Gly Trp Ser Asp Ser Glu Ser Glu Lys Leu Leu Ser Arg		
	35	40
Ser Met Ser Asn Asp Ser Ser Lys Thr Trp Asp Ala Val Ser Glu Asp		
	50	55
Ser Val Phe Asp Pro Asp Gly Ser Pro Leu Leu Lys Asp Arg Leu Gly		
65	70	75
Asn Leu Asp Phe Lys Tyr Ile Glu Arg Asp Pro Pro His Lys Arg Ile		
	85	90
Pro Leu Thr Asp Lys Ile Asn Val Leu Val Glu Lys Tyr Pro Gly Leu		
	100	105
Met Thr Leu Arg Ser Val Asp Met Ser Pro Ala Ser Trp Met Ala Val		
	115	120
Ala Trp Tyr Pro Ile Tyr His Ile Pro Thr Cys Arg Asn Glu Lys Asp		
	130	135
Leu Thr Thr Gly Phe Leu Thr Tyr His Thr Leu Ser Ser Ser Phe Gln		
145	150	155
Asp Asn Val Val Glu Gly Asp Gln Ser Asn Asn Asn Glu Glu Thr Glu		
	165	170
Phe Cys Glu Asp Ser Val Ile Asn Lys Arg Met Pro Leu Pro Pro Phe		
	180	185
Gly Val Thr Thr Tyr Lys Met Gln Gly Asp Leu Trp Gly Lys Thr Gly		
	195	200
Phe Asp Gln Asp Arg Leu Leu Tyr Leu Gln Ser Ala Ala Asp Ser Trp		
210	215	220
Leu Lys Gln Leu Asn Val Asp His His Asp Tyr Asn Phe Phe Leu Asn		
225	230	235
Ser Ser Phe		240

(2) INFORMATION FOR SEQ ID NO:88:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 1235 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..1235

(D) OTHER INFORMATION: / Ceres Seq. ID 1481621

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:88:

```
attgacccaa cgcttcttct ccggcagcagc tgttcagagt tcgattttcca ttttcggggtc      60
gaaagggttga ttttatttgg attttggatg gtagattcag ttcctaaaca taggaaactt      120
gaatctcaga gtttttcgag ttagggataa gagaaagaaa cacagttgga gttatactga      180
tgaatggagg aggctcgagt agtttgcgtt cagcattgtc ctattgtgtg cagcaagtac      240
gaaactatga ctatcatcac tacctctgtc tccttgaact cccaactgag atgcgtaaag      300
cagcatttgc tctccgggct tttaatgtag aaaccgcaag agccatggat gttgcatctg      360
atcccaaaat cggcttgatg cggttacttt ggtggcaaga agcaattgac aaactctaca      420
ccaaaaagcc cataaaccat ccagctgcac aagctctgtc ttgggcaata tcagaacata      480
acatcagtaa acccttggcta aaacgctcgg ttgacgctag aatccgagat gcccaaagag      540
aagtagacga tataccagag agcattgcgg agctagagaa atacgcagaa gacacagttt      600
ccactcttct gtacaataca ctccaagcag gcggaattag ttcaacaaca gctgatcatg      660
cagcttcaca catttggtaaa gccagtggtc ttgtcttgct gcttaaatca ttaccgtacc      720
actgtaccag aaaccgtcac cagagttaca tccctgcaga tctcgtgag aagcacgggt      780
tgctcgtgaa acaaggtgga cgattagaaa ttcttctgga taacgattca agagaaggac      840
taagcaatgt cgtgtttgag attgcatctg ttgccaatgc acatctcctg aaagcccgtg      900
aactggcggg aaaggttcct gcagaagcta aaccggtagt gcttcattct gtgccggtac      960
aagttcttct ggattcggtt aataaagtac aattcgatgt gtttgatccc aggattcaaa      1020
gaggagttct tgggtttcct ccactcttgt ttcagtttaa actcaagtgg tattcatgga      1080
gagcaatgtt ttgaaaactt gtctttatct cccttttctt gcctctttta tttctgggtt      1140
caaagacttt acattaaact ccagcttact tgtatttctt ttgtaataat acaaaaattac      1200
aaatggtgat gaatacaaaa taaagaattt gtttc
```

(2) INFORMATION FOR SEQ ID NO:89:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 304 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..304

(D) OTHER INFORMATION: / Ceres Seq. ID 1481622

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:89:

```
Met Asn Gly Gly Ser Ser Ser Leu Arg Ser Ala Leu Ser Tyr Cys
1      5      10      15
Val Gln Gln Val Arg Asn Tyr Asp Tyr His His Tyr Leu Cys Leu Leu
20     25     30
Glu Leu Pro Thr Glu Met Arg Lys Ala Ala Phe Ala Leu Arg Ala Phe
35     40     45
Asn Val Glu Thr Ala Arg Ala Met Asp Val Ala Ser Asp Pro Lys Ile
50     55     60
Gly Leu Met Arg Leu Leu Trp Trp Gln Glu Ala Ile Asp Lys Leu Tyr
65     70     75     80
Thr Lys Lys Pro Ile Asn His Pro Ala Ala Gln Ala Leu Ser Trp Ala
85     90     95
Ile Ser Glu His Asn Ile Ser Lys Pro Trp Leu Lys Arg Ser Val Asp
100    105    110
Ala Arg Ile Arg Asp Ala Gln Arg Glu Val Asp Asp Ile Pro Glu Ser
115    120    125
Ile Ala Glu Leu Glu Lys Tyr Ala Glu Asp Thr Val Ser Thr Leu Leu
130    135    140
Tyr Asn Thr Leu Gln Ala Gly Gly Ile Ser Ser Thr Thr Ala Asp His
145    150    155    160
```

Ala	Ala	Ser	His	Ile	Gly	Lys	Ala	Ser	Gly	Leu	Val	Leu	Leu	Leu	Lys
				165					170					175	
Ser	Leu	Pro	Tyr	His	Cys	Thr	Arg	Asn	Arg	His	Gln	Ser	Tyr	Ile	Pro
			180					185					190		
Ala	Asp	Leu	Ala	Glu	Lys	His	Gly	Leu	Leu	Val	Lys	Gln	Gly	Gly	Arg
		195					200				205				
Leu	Glu	Ile	Leu	Leu	Asp	Asn	Asp	Ser	Arg	Glu	Gly	Leu	Ser	Asn	Val
	210					215					220				
Val	Phe	Glu	Ile	Ala	Ser	Val	Ala	Asn	Ala	His	Leu	Leu	Lys	Ala	Arg
225					230					235				240	
Glu	Leu	Ala	Gly	Lys	Val	Pro	Ala	Glu	Ala	Lys	Pro	Val	Leu	Leu	His
				245					250					255	
Ser	Val	Pro	Val	Gln	Val	Leu	Leu	Asp	Ser	Leu	Asn	Lys	Val	Gln	Phe
			260					265				270			
Asp	Val	Phe	Asp	Pro	Arg	Ile	Gln	Arg	Gly	Val	Leu	Gly	Val	Pro	Pro
		275					280					285			
Leu	Leu	Phe	Gln	Phe	Lys	Leu	Lys	Trp	Tyr	Ser	Trp	Arg	Ala	Met	Phe
	290					295					300				

(2) INFORMATION FOR SEQ ID NO:90:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 267 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..267

(D) OTHER INFORMATION: / Ceres Seq. ID 1481623

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:90:

Met	Arg	Lys	Ala	Ala	Phe	Ala	Leu	Arg	Ala	Phe	Asn	Val	Glu	Thr	Ala
1			5						10					15	
Arg	Ala	Met	Asp	Val	Ala	Ser	Asp	Pro	Lys	Ile	Gly	Leu	Met	Arg	Leu
			20					25					30		
Leu	Trp	Trp	Gln	Glu	Ala	Ile	Asp	Lys	Leu	Tyr	Thr	Lys	Lys	Pro	Ile
			35				40					45			
Asn	His	Pro	Ala	Ala	Gln	Ala	Leu	Ser	Trp	Ala	Ile	Ser	Glu	His	Asn
	50					55				60					
Ile	Ser	Lys	Pro	Trp	Leu	Lys	Arg	Ser	Val	Asp	Ala	Arg	Ile	Arg	Asp
65					70					75				80	
Ala	Gln	Arg	Glu	Val	Asp	Asp	Ile	Pro	Glu	Ser	Ile	Ala	Glu	Leu	Glu
				85					90					95	
Lys	Tyr	Ala	Glu	Asp	Thr	Val	Ser	Thr	Leu	Leu	Tyr	Asn	Thr	Leu	Gln
			100					105					110		
Ala	Gly	Gly	Ile	Ser	Ser	Thr	Thr	Ala	Asp	His	Ala	Ala	Ser	His	Ile
	115						120						125		
Gly	Lys	Ala	Ser	Gly	Leu	Val	Leu	Leu	Lys	Ser	Leu	Pro	Tyr	His	
	130					135				140					
Cys	Thr	Arg	Asn	Arg	His	Gln	Ser	Tyr	Ile	Pro	Ala	Asp	Leu	Ala	Glu
145					150					155				160	
Lys	His	Gly	Leu	Leu	Val	Lys	Gln	Gly	Gly	Arg	Leu	Glu	Ile	Leu	Leu
				165					170					175	
Asp	Asn	Asp	Ser	Arg	Glu	Gly	Leu	Ser	Asn	Val	Val	Phe	Glu	Ile	Ala
			180					185					190		
Ser	Val	Ala	Asn	Ala	His	Leu	Leu	Lys	Ala	Arg	Glu	Leu	Ala	Gly	Lys
		195					200					205			
Val	Pro	Ala	Glu	Ala	Lys	Pro	Val	Leu	Leu	His	Ser	Val	Pro	Val	Gln

210	215	220
Val Leu Leu Asp Ser Leu Asn Lys Val Gln Phe Asp Val Phe Asp Pro		
225	230	235
Arg Ile Gln Arg Gly Val Leu Gly Val Pro Pro Leu Leu Phe Gln Phe		240
	245	250
		255
Lys Leu Lys Trp Tyr Ser Trp Arg Ala Met Phe		
260	265	

(2) INFORMATION FOR SEQ ID NO:91:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 249 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..249

(D) OTHER INFORMATION: / Ceres Seq. ID 1481624

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:91:

Met Asp Val Ala Ser Asp Pro Lys Ile Gly Leu Met Arg Leu Leu Trp		
1	5	10
Trp Gln Glu Ala Ile Asp Lys Leu Tyr Thr Lys Lys Pro Ile Asn His		15
	20	25
Pro Ala Ala Gln Ala Leu Ser Trp Ala Ile Ser Glu His Asn Ile Ser		30
	35	40
Lys Pro Trp Leu Lys Arg Ser Val Asp Ala Arg Ile Arg Asp Ala Gln		45
	50	55
Arg Glu Val Asp Asp Ile Pro Glu Ser Ile Ala Glu Leu Glu Lys Tyr		60
	65	70
Ala Glu Asp Thr Val Ser Thr Leu Leu Tyr Asn Thr Leu Gln Ala Gly		75
	85	90
Gly Ile Ser Ser Thr Thr Ala Asp His Ala Ala Ser His Ile Gly Lys		95
	100	105
Ala Ser Gly Leu Val Leu Leu Leu Lys Ser Leu Pro Tyr His Cys Thr		110
	115	120
Arg Asn Arg His Gln Ser Tyr Ile Pro Ala Asp Leu Ala Glu Lys His		125
	130	135
Gly Leu Leu Val Lys Gln Gly Gly Arg Leu Glu Ile Leu Leu Asp Asn		140
	145	150
Asp Ser Arg Glu Gly Leu Ser Asn Val Val Phe Glu Ile Ala Ser Val		155
	165	170
Ala Asn Ala His Leu Leu Lys Ala Arg Glu Leu Ala Gly Lys Val Pro		175
	180	185
Ala Glu Ala Lys Pro Val Leu Leu His Ser Val Pro Val Gln Val Leu		190
	195	200
Leu Asp Ser Leu Asn Lys Val Gln Phe Asp Val Phe Asp Pro Arg Ile		205
	210	215
Gln Arg Gly Val Leu Gly Val Pro Pro Leu Leu Phe Gln Phe Lys Leu		220
	225	230
Lys Trp Tyr Ser Trp Arg Ala Met Phe		235
	245	240

(2) INFORMATION FOR SEQ ID NO:92:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 1232 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -
(B) LOCATION: 1..1232
(D) OTHER INFORMATION: / Ceres Seq. ID 1481625

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:92:

ctctctcgcg	ttcgtttctta	tccacgagct	ctgcaccgtc	gcaatctccg	tcttctccat	60
ttagatccaa	cacagagcct	tttctacatg	aaattcggca	aagagtttcg	tactcacctc	120
gaagaaactt	taccagagtg	gagagacaag	ttcctttgct	ataaaccttt	aaaaaagctt	180
ctcaaatatt	atccttatta	ctccgccgat	tttggaccgc	ccaattccga	tcacaacgat	240
tcgcgtccag	tatttgctga	tactactaac	atctcttccg	ccgccgacga	cggcggtgtg	300
gtcccgcgcg	tcaggccatc	ggaagatctc	caggggttcgt	ttgtgaggat	acttaacgat	360
gaacttgaga	agtttaacga	tttttacgtt	gataaggaag	aagatttcgt	tatcagatta	420
caggagctca	aggaaagaat	cgagcaagtt	aaagaaaaga	atggggaatt	tgcatacaga	480
agtgaattca	gcgaagaaat	gatggatatt	cggagagacc	ttgttaccat	tcattggcag	540
atggtgctcc	tgaaaaacta	cagctccctt	aattttgcag	gacttgtcaa	gattttgaag	600
aagtacgata	aaagaacagg	tggactttta	cgtttgcctt	tcacacagct	tgttctccat	660
caacccttct	ttactacaga	gccacttact	aggttagtcc	gtgaatgtga	ggccaatctt	720
gagcttcttt	ttccttcaga	agcgggaagt	gtagagtctt	ctagcgcagt	gcaagcacac	780
tcaagctcac	atcagcacia	ctccccaaga	atctcagctg	agacttcctc	aactctcggc	840
aatgaaaatc	ttgatataata	taagagtaca	ctcgctgcaa	tgagagctat	aagagggtta	900
caaaaggcta	gctcgacgta	caacccttta	tcattctcat	cgcttcttca	gaacgaggat	960
gatgagacgg	taacagctga	aaactctcca	aactctggga	acaaagatga	ttcagagaag	1020
gaagatactg	gaccttccca	ctgatacaga	gagaatgatg	ctctttttga	tcaagatttt	1080
gagaatttgc	ttcttgattt	caccttaact	tttcataaaa	ttaacacatt	ttactttact	1140
tcttcacctt	ttgcaggaca	caacttctgt	atgcatttga	atttttagtac	agtcgtttat	1200
agattttcaa	tgaaattttc	ctccattgtc	gc			

(2) INFORMATION FOR SEQ ID NO:93:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 347 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide
(B) LOCATION: 1..347
(D) OTHER INFORMATION: / Ceres Seq. ID 1481626

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:93:

Leu	Ser	Arg	Val	Arg	Ser	Tyr	Pro	Arg	Ala	Leu	His	Arg	Arg	Asn	Leu	
1			5					10						15		
Arg	Leu	Leu	His	Leu	Asp	Pro	Thr	Gln	Ser	Leu	Phe	Tyr	Met	Lys	Phe	
			20					25					30			
Gly	Lys	Glu	Phe	Arg	Thr	His	Leu	Glu	Glu	Thr	Leu	Pro	Glu	Trp	Arg	
			35				40					45				
Asp	Lys	Phe	Leu	Cys	Tyr	Lys	Pro	Leu	Lys	Lys	Leu	Leu	Lys	Tyr	Tyr	
			50			55					60					
Pro	Tyr	Tyr	Ser	Ala	Asp	Phe	Gly	Pro	Ala	Asn	Ser	Asp	His	Asn	Asp	
65				70						75				80		
Ser	Arg	Pro	Val	Phe	Ala	Asp	Thr	Thr	Asn	Ile	Ser	Ser	Ala	Ala	Asp	
			85					90						95		
Asp	Gly	Gly	Val	Val	Pro	Gly	Val	Arg	Pro	Ser	Glu	Asp	Leu	Gln	Gly	
			100				105						110			
Ser	Phe	Val	Arg	Ile	Leu	Asn	Asp	Glu	Leu	Glu	Lys	Phe	Asn	Asp	Phe	
			115			120						125				
Tyr	Val	Asp	Lys	Glu	Glu	Asp	Phe	Val	Ile	Arg	Leu	Gln	Glu	Leu	Lys	
			130			135					140					
Glu	Arg	Ile	Glu	Gln	Val	Lys	Glu	Lys	Asn	Gly	Glu	Phe	Ala	Ser	Glu	
145				150					155					160		
Ser	Glu	Phe	Ser	Glu	Glu	Met	Met	Asp	Ile	Arg	Arg	Asp	Leu	Val	Thr	
			165					170						175		

Ile His Gly Glu Met Val Leu Leu Lys Asn Tyr Ser Ser Leu Asn Phe
180 185 190
Ala Gly Leu Val Lys Ile Leu Lys Lys Tyr Asp Lys Arg Thr Gly Gly
195 200 205
Leu Leu Arg Leu Pro Phe Thr Gln Leu Val Leu His Gln Pro Phe Phe
210 215 220
Thr Thr Glu Pro Leu Thr Arg Leu Val Arg Glu Cys Glu Ala Asn Leu
225 230 235 240
Glu Leu Leu Phe Pro Ser Glu Ala Glu Val Val Glu Ser Ser Ser Ala
245 250 255
Val Gln Ala His Ser Ser Ser His Gln His Asn Ser Pro Arg Ile Ser
260 265 270
Ala Glu Thr Ser Ser Thr Leu Gly Asn Glu Asn Leu Asp Ile Tyr Lys
275 280 285
Ser Thr Leu Ala Ala Met Arg Ala Ile Arg Gly Leu Gln Lys Ala Ser
290 295 300
Ser Thr Tyr Asn Pro Leu Ser Phe Ser Ser Leu Leu Gln Asn Glu Asp
305 310 315 320
Asp Glu Thr Val Thr Ala Glu Asn Ser Pro Asn Ser Gly Asn Lys Asp
325 330 335
Asp Ser Glu Lys Glu Asp Thr Gly Pro Ser His
340 345

(2) INFORMATION FOR SEQ ID NO:94:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 318 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..318

(D) OTHER INFORMATION: / Ceres Seq. ID 1481627

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:94:

Met Lys Phe Gly Lys Glu Phe Arg Thr His Leu Glu Glu Thr Leu Pro
1 5 10 15
Glu Trp Arg Asp Lys Phe Leu Cys Tyr Lys Pro Leu Lys Lys Leu Leu
20 25 30
Lys Tyr Tyr Pro Tyr Tyr Ser Ala Asp Phe Gly Pro Ala Asn Ser Asp
35 40 45
His Asn Asp Ser Arg Pro Val Phe Ala Asp Thr Thr Asn Ile Ser Ser
50 55 60
Ala Ala Asp Asp Gly Gly Val Val Pro Gly Val Arg Pro Ser Glu Asp
65 70 75 80
Leu Gln Gly Ser Phe Val Arg Ile Leu Asn Asp Glu Leu Glu Lys Phe
85 90 95
Asn Asp Phe Tyr Val Asp Lys Glu Glu Asp Phe Val Ile Arg Leu Gln
100 105 110
Glu Leu Lys Glu Arg Ile Glu Gln Val Lys Glu Lys Asn Gly Glu Phe
115 120 125
Ala Ser Glu Ser Glu Phe Ser Glu Glu Met Met Asp Ile Arg Arg Asp
130 135 140
Leu Val Thr Ile His Gly Glu Met Val Leu Leu Lys Asn Tyr Ser Ser
145 150 155 160
Leu Asn Phe Ala Gly Leu Val Lys Ile Leu Lys Lys Tyr Asp Lys Arg
165 170 175
Thr Gly Gly Leu Leu Arg Leu Pro Phe Thr Gln Leu Val Leu His Gln
180 185 190
Pro Phe Phe Thr Thr Glu Pro Leu Thr Arg Leu Val Arg Glu Cys Glu

195	200	205
Ala Asn Leu Glu Leu Leu Phe Pro Ser Glu Ala Glu Val Val Glu Ser		
210	215	220
Ser Ser Ala Val Gln Ala His Ser Ser Ser His Gln His Asn Ser Pro		
225	230	235
Arg Ile Ser Ala Glu Thr Ser Ser Thr Leu Gly Asn Glu Asn Leu Asp		
245	250	255
Ile Tyr Lys Ser Thr Leu Ala Ala Met Arg Ala Ile Arg Gly Leu Gln		
260	265	270
Lys Ala Ser Ser Thr Tyr Asn Pro Leu Ser Phe Ser Ser Leu Leu Gln		
275	280	285
Asn Glu Asp Asp Glu Thr Val Thr Ala Glu Asn Ser Pro Asn Ser Gly		
290	295	300
Asn Lys Asp Asp Ser Glu Lys Glu Asp Thr Gly Pro Ser His		
305	310	315

(2) INFORMATION FOR SEQ ID NO:95:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 181 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..181
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481628

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:95:

Met Met Asp Ile Arg Arg Asp Leu Val Thr Ile His Gly Glu Met Val		
1	5	10
Leu Leu Lys Asn Tyr Ser Ser Leu Asn Phe Ala Gly Leu Val Lys Ile		
20	25	30
Leu Lys Lys Tyr Asp Lys Arg Thr Gly Gly Leu Leu Arg Leu Pro Phe		
35	40	45
Thr Gln Leu Val Leu His Gln Pro Phe Phe Thr Thr Glu Pro Leu Thr		
50	55	60
Arg Leu Val Arg Glu Cys Glu Ala Asn Leu Glu Leu Leu Phe Pro Ser		
65	70	75
Glu Ala Glu Val Val Glu Ser Ser Ser Ala Val Gln Ala His Ser Ser		
85	90	95
Ser His Gln His Asn Ser Pro Arg Ile Ser Ala Glu Thr Ser Ser Thr		
100	105	110
Leu Gly Asn Glu Asn Leu Asp Ile Tyr Lys Ser Thr Leu Ala Ala Met		
115	120	125
Arg Ala Ile Arg Gly Leu Gln Lys Ala Ser Ser Thr Tyr Asn Pro Leu		
130	135	140
Ser Phe Ser Ser Leu Leu Gln Asn Glu Asp Asp Glu Thr Val Thr Ala		
145	150	155
Glu Asn Ser Pro Asn Ser Gly Asn Lys Asp Asp Ser Glu Lys Glu Asp		
165	170	175
Thr Gly Pro Ser His		
180		

(2) INFORMATION FOR SEQ ID NO:96:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1217 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..1217

(D) OTHER INFORMATION: / Ceres Seq. ID 1481632

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:96:

actccaaaat	gaggacgcgc	cggcaaacat	atccgccgat	cgctgaatcc	ctcacggcga	60
gggccattgt	tcaggcgctt	cggcgctcag	ccacaatatc	aggaaatggt	ggaccaaaga	120
agaagaagaa	ctgtgttaat	agaggattgt	gggataaaca	gattccgacg	gatctgctgc	180
aagagatact	gtcttgccctc	ggattaaaag	ccaacataca	tgcttctctc	gtctgcaaga	240
catggcttaa	agaagctggt	tctgtcagga	agtttcagag	tcgtccttgg	cttttttatc	300
cacagagtca	gagaggagga	ccaaaagaag	gagactacgt	tctctttaac	ccatcacggt	360
ctcaaacaca	tcacctcaag	tttccagagt	taacgggcta	cagaaataaa	ttagcttgtg	420
ctaaggatgg	ttggttgctt	gtggtaaaag	ataaccccg	tgtggtcttc	tttcttaacc	480
cgtttaccgg	ggaacgcac	tgcttacccc	aggtgccaca	aaattccaca	cgcgattgct	540
taactttctc	agccgctccc	acatcaacta	gttgttgctg	catatccttc	acccctcaaa	600
gtttttctta	cgcagttggt	aaagttgata	cttggcgccc	tgggtgaatcc	gtatggacca	660
ctcatcactt	tgatcaaaa	cgttacgggt	aggtaatcaa	tagatgtatc	ttctccaatg	720
gtatgttcta	ttgtctcagt	accagtggcc	gcctctcggt	tttcgaccgc	tctagagaaa	780
cctggaatgt	tcttccagtg	aaacatgtc	gggcctttcg	tcgtaaaatt	atgcttgtga	840
ggcaagtatt	catgacagag	catgaaggag	acatctttgt	tgtgactaca	cgccgcgtaa	900
acaacagaaa	actggttgcc	tttaactaaa	accttcaagg	caatgtgtgg	gaagagatga	960
aagtacctaa	tggttgaca	gtattttcaa	gtgacgctac	ctctttaaca	agagctggtc	1020
ttccagagga	ggagaggaac	attctatatt	catcggatat	cgatgatttt	gtgaaaagct	1080
ctcatccaac	tttctattat	tatgactgca	gcgcttggct	ccagccacct	catgacaatt	1140
ttaatttttg	actatcatcc	ttaagtgttt	ttgtttttga	aaaaacatgt	tttaataacct	1200
tttaaaagctt	ttgatttc					

(2) INFORMATION FOR SEQ ID NO:97:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 382 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..382

(D) OTHER INFORMATION: / Ceres Seq. ID 1481633

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:97:

Ser	Lys	Met	Arg	Thr	Arg	Arg	Gln	Thr	Tyr	Pro	Pro	Ile	Ala	Glu	Ser
1				5					10					15	
Leu	Thr	Ala	Arg	Ser	Ile	Val	Gln	Ala	Leu	Pro	Ala	Ser	Ala	Thr	Ile
			20					25					30		
Ser	Gly	Asn	Gly	Gly	Pro	Lys	Lys	Lys	Lys	Asn	Cys	Val	Asn	Arg	Gly
		35					40					45			
Leu	Trp	Asp	Lys	Gln	Ile	Pro	Thr	Asp	Leu	Leu	Gln	Glu	Ile	Leu	Ser
		50				55					60				
Cys	Leu	Gly	Leu	Lys	Ala	Asn	Ile	His	Ala	Ser	Leu	Val	Cys	Lys	Thr
65					70					75				80	
Trp	Leu	Lys	Glu	Ala	Val	Ser	Val	Arg	Lys	Phe	Gln	Ser	Arg	Pro	Trp
			85						90					95	
Leu	Phe	Tyr	Pro	Gln	Ser	Gln	Arg	Gly	Gly	Pro	Lys	Glu	Gly	Asp	Tyr
			100					105					110		
Val	Leu	Phe	Asn	Pro	Ser	Arg	Ser	Gln	Thr	His	His	Leu	Lys	Phe	Pro
		115					120					125			
Glu	Leu	Thr	Gly	Tyr	Arg	Asn	Lys	Leu	Ala	Cys	Ala	Lys	Asp	Gly	Trp
		130					135				140				
Leu	Leu	Val	Val	Lys	Asp	Asn	Pro	Asp	Val	Val	Phe	Phe	Leu	Asn	Pro
145					150					155				160	
Phe	Thr	Gly	Glu	Arg	Ile	Cys	Leu	Pro	Gln	Val	Pro	Gln	Asn	Ser	Thr
			165					170						175	

Arg Asp Cys Leu Thr Phe Ser Ala Ala Pro Thr Ser Thr Ser Cys Cys
180 185 190
Val Ile Ser Phe Thr Pro Gln Ser Phe Leu Tyr Ala Val Val Lys Val
195 200 205
Asp Thr Trp Arg Pro Gly Glu Ser Val Trp Thr Thr His His Phe Asp
210 215 220
Gln Lys Arg Tyr Gly Glu Val Ile Asn Arg Cys Ile Phe Ser Asn Gly
225 230 235 240
Met Phe Tyr Cys Leu Ser Thr Ser Gly Arg Leu Ser Phe Phe Asp Pro
245 250 255
Ser Arg Glu Thr Trp Asn Val Leu Pro Val Lys Pro Cys Arg Ala Phe
260 265 270
Arg Arg Lys Ile Met Leu Val Arg Gln Val Phe Met Thr Glu His Glu
275 280 285
Gly Asp Ile Phe Val Val Thr Thr Arg Arg Val Asn Asn Arg Lys Leu
290 295 300
Leu Ala Phe Lys Leu Asn Leu Gln Gly Asn Val Trp Glu Glu Met Lys
305 310 315 320
Val Pro Asn Gly Leu Thr Val Phe Ser Ser Asp Ala Thr Ser Leu Thr
325 330 335
Arg Ala Gly Leu Pro Glu Glu Glu Arg Asn Ile Leu Tyr Ser Ser Asp
340 345 350
Ile Asp Asp Phe Val Lys Ser Ser His Pro Thr Phe Tyr Tyr Tyr Asp
355 360 365
Cys Ser Ala Trp Leu Gln Pro Pro His Asp Asn Phe Asn Phe
370 375 380

(2) INFORMATION FOR SEQ ID NO:98:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 380 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..380
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481634

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:98:

Met Arg Thr Arg Arg Gln Thr Tyr Pro Pro Ile Ala Glu Ser Leu Thr
1 5 10 15
Ala Arg Ser Ile Val Gln Ala Leu Pro Ala Ser Ala Thr Ile Ser Gly
20 25 30
Asn Gly Gly Pro Lys Lys Lys Lys Asn Cys Val Asn Arg Gly Leu Trp
35 40 45
Asp Lys Gln Ile Pro Thr Asp Leu Leu Gln Glu Ile Leu Ser Cys Leu
50 55 60
Gly Leu Lys Ala Asn Ile His Ala Ser Leu Val Cys Lys Thr Trp Leu
65 70 75 80
Lys Glu Ala Val Ser Val Arg Lys Phe Gln Ser Arg Pro Trp Leu Phe
85 90 95
Tyr Pro Gln Ser Gln Arg Gly Gly Pro Lys Glu Gly Asp Tyr Val Leu
100 105 110
Phe Asn Pro Ser Arg Ser Gln Thr His His Leu Lys Phe Pro Glu Leu
115 120 125
Thr Gly Tyr Arg Asn Lys Leu Ala Cys Ala Lys Asp Gly Trp Leu Leu
130 135 140
Val Val Lys Asp Asn Pro Asp Val Val Phe Phe Leu Asn Pro Phe Thr
145 150 155 160
Gly Glu Arg Ile Cys Leu Pro Gln Val Pro Gln Asn Ser Thr Arg Asp

				165					170					175				
Cys	Leu	Thr	Phe	Ser	Ala	Ala	Pro	Thr	Ser	Thr	Ser	Cys	Cys	Val	Ile			
			180						185				190					
Ser	Phe	Thr	Pro	Gln	Ser	Phe	Leu	Tyr	Ala	Val	Val	Lys	Val	Asp	Thr			
		195					200					205						
Trp	Arg	Pro	Gly	Glu	Ser	Val	Trp	Thr	Thr	His	His	Phe	Asp	Gln	Lys			
	210					215					220							
Arg	Tyr	Gly	Glu	Val	Ile	Asn	Arg	Cys	Ile	Phe	Ser	Asn	Gly	Met	Phe			
	225				230					235					240			
Tyr	Cys	Leu	Ser	Thr	Ser	Gly	Arg	Leu	Ser	Phe	Phe	Asp	Pro	Ser	Arg			
				245					250					255				
Glu	Thr	Trp	Asn	Val	Leu	Pro	Val	Lys	Pro	Cys	Arg	Ala	Phe	Arg	Arg			
			260					265					270					
Lys	Ile	Met	Leu	Val	Arg	Gln	Val	Phe	Met	Thr	Glu	His	Glu	Gly	Asp			
		275					280					285						
Ile	Phe	Val	Val	Thr	Thr	Arg	Arg	Val	Asn	Asn	Arg	Lys	Leu	Leu	Ala			
	290					295					300							
Phe	Lys	Leu	Asn	Leu	Gln	Gly	Asn	Val	Trp	Glu	Glu	Met	Lys	Val	Pro			
	305				310					315					320			
Asn	Gly	Leu	Thr	Val	Phe	Ser	Ser	Asp	Ala	Thr	Ser	Leu	Thr	Arg	Ala			
				325					330					335				
Gly	Leu	Pro	Glu	Glu	Glu	Arg	Asn	Ile	Leu	Tyr	Ser	Ser	Asp	Ile	Asp			
		340						345					350					
Asp	Phe	Val	Lys	Ser	Ser	His	Pro	Thr	Phe	Tyr	Tyr	Tyr	Asp	Cys	Ser			
		355					360					365						
Ala	Trp	Leu	Gln	Pro	Pro	His	Asp	Asn	Phe	Asn	Phe							
	370					375					380							

(2) INFORMATION FOR SEQ ID NO:99:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 667 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..667
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481635

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:99:

mtgacgcttg	gysttcttgc	accctaccgt	gggattggca	ttggtaagtc	ataaatctgc	60
tgtgccaac	ttgaataccg	acaaaacatg	aataatgaag	aaaactgtta	gcatgttgtg	120
taattctgtt	gttttcctta	cgttttaaca	catgaggcca	gctgtagtat	gtttattctg	180
tagttctcta	tttgaagtgt	ctccatttag	agattcaaac	caccaagaaa	tagtccttag	240
ggttttatgc	atatcggtgt	tttaccgaga	aactggaatt	agtgactatg	atttcctcct	300
atatcaagat	ttaagatcga	attccctgct	tttagaaaaga	aaaactcgat	gtctataatt	360
tgtgtatcct	gtttttttcg	tcttttgacg	gtc aaaatct	attgaatcat	gttccttgaca	420
tgtgctccaa	gcaaaacatg	tgtgagatat	acttgcatgt	gcagacaaac	aacgaagacg	480
caatcaagtt	ctacaagaag	ttcggctttg	agatcacaga	taccatacaa	aactattaca	540
tcaacattga	gccaagagat	tgctacgttg	tcagcaagtc	ctttgctcaa	tctgaagcca	600
acaaatgatg	aaaaatacca	aacttgggga	agrcattcct	ccccagtttc	tttggttgc	660
tcagtttc						

(2) INFORMATION FOR SEQ ID NO:100:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 68 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide
(B) LOCATION: 1..68
(D) OTHER INFORMATION: / Ceres Seq. ID 1481636

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:100:

Met Arg Pro Ala Val Val Cys Leu Phe Cys Ser Ser Leu Phe Glu Val
1 5 10 15
Ser Pro Phe Arg Asp Ser Asn His Gln Glu Ile Val Leu Arg Val Leu
20 25 30
Cys Ile Ser Leu Phe Tyr Arg Glu Thr Gly Ile Ser Asp Tyr Asp Phe
35 40 45
Leu Leu Tyr Gln Asp Leu Arg Ser Asn Ser Leu Leu Leu Glu Arg Lys
50 55 60
Thr Arg Cys Leu
65

(2) INFORMATION FOR SEQ ID NO:101:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 86 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide
(B) LOCATION: 1..86
(D) OTHER INFORMATION: / Ceres Seq. ID 1481637

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:101:

Met Phe Leu Thr Cys Ala Pro Ser Lys Thr Cys Val Arg Tyr Thr Cys
1 5 10 15
Met Cys Arg Gln Thr Thr Lys Thr Gln Ser Ser Ser Thr Arg Ser Ser
20 25 30
Ala Leu Arg Ser Gln Ile Pro Tyr Lys Thr Ile Thr Ser Thr Leu Ser
35 40 45
Gln Glu Ile Ala Thr Leu Ser Ala Ser Pro Leu Leu Asn Leu Lys Pro
50 55 60
Thr Asn Asp Glu Lys Tyr Gln Thr Trp Gly Ser His Ser Ser Pro Val
65 70 75 80
Ser Leu Leu His Ser Val
85

(2) INFORMATION FOR SEQ ID NO:102:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 70 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide
(B) LOCATION: 1..70
(D) OTHER INFORMATION: / Ceres Seq. ID 1481638

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:102:

Met Cys Arg Gln Thr Thr Lys Thr Gln Ser Ser Ser Thr Arg Ser Ser
1 5 10 15
Ala Leu Arg Ser Gln Ile Pro Tyr Lys Thr Ile Thr Ser Thr Leu Ser
20 25 30
Gln Glu Ile Ala Thr Leu Ser Ala Ser Pro Leu Leu Asn Leu Lys Pro
35 40 45
Thr Asn Asp Glu Lys Tyr Gln Thr Trp Gly Ser His Ser Ser Pro Val
50 55 60
Ser Leu Leu His Ser Val

65

70

(2) INFORMATION FOR SEQ ID NO:103:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1177 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1177
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481639

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:103:

```
ctttgcgatg gtaaaatagt gttctcagat tgggctagca atgtgagttc tattctttgg      60
gtccctgttc gtgccttgag tgagaagcct gctagagggt catcatcagt cactccgctg      120
aaacaagata ttttagaggg aatgagaact gtagctttga aacttgaatt tggggttcat      180
cataaccaga tatttgagag aaccatagct gcacatttta ccgatccctt tgatgtgacc      240
acaagggtgg caaacaatg caatgatggc actttggctt tgcaggttat gttacactcc      300
ctcgtcaagg cgaacttgat agttcttgat gtttggcttg atcttcaaga tggattttatt      360
catggacaaa atgatggaag accgacttca acgttctttc cgcttgctcg gtctccagga      420
tctagagcag cagtcgtggt cagtatatgc ctagacaaga gtatgtcatc agaagggaaa      480
gatttgcagc taccagaaag cattctgaat atcaaatatg gaatccatgg ggatagagca      540
gctggagcac acaggccagt ggatgcagat cactctgaaa ctgatactta agggagagat      600
ttggtgttca agagtgcctat tgttttgcag cgtccagtac ttgatccttg cctcacagtt      660
ggattcctcc cacttccttc tgatgggctt agggtcggga aacttatcac catgcagtgg      720
agagtggaaa ggcttaaaga tctcaaagaa agtgaagccg tggaaacaaca acatgatgag      780
gtgttatatg aagtcaatgc aaattcggag aattggatga tcgctggtag gaagagaggc      840
catgtctctc tctcagagga gcaaggttca agagtagtaa tctcgatact atgtgtcccg      900
ttagtgcggt gttatgtccg tctcctcctc ctcgggttgc caaacgtaga agaagcaaat      960
gtaagcagca atccatcggg tctcacttta gtatgtgtct tgcctccact tctcagttct      1020
tctactgcgt tacctgtcaa gtaatagaat ctcactctat attttttcca agaaaacatt      1080
ttttctgtat ttttattttg tttgcgatca aagaaatc atc agagtatggg atcatcaatg      1140
atgagagtga tttttctttt gtgacgattt tatttcc
```

(2) INFORMATION FOR SEQ ID NO:104:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 196 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..196
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481640

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:104:

```
Leu Cys Asp Gly Lys Ile Val Phe Ser Asp Trp Ala Ser Asn Val Ser
1          5          10          15
Ser Ile Leu Trp Val Pro Val Arg Ala Leu Ser Glu Lys Leu Ala Arg
20          25          30
Gly Ser Ser Ser Val Thr Pro Leu Lys Gln Asp Ile Leu Glu Gly Met
35          40          45
Arg Thr Val Ala Leu Lys Leu Glu Phe Gly Val His His Asn Gln Ile
50          55          60
Phe Glu Arg Thr Ile Ala Ala His Phe Thr Asp Pro Phe Asp Val Thr
65          70          75          80
Thr Arg Val Ala Asn Lys Cys Asn Asp Gly Thr Leu Val Leu Gln Val
85          90          95
Met Leu His Ser Leu Val Lys Ala Asn Leu Ile Val Leu Asp Val Trp
100          105          110
```

Leu Asp Leu Gln Asp Gly Phe Ile His Gly Gln Asn Asp Gly Arg Pro
115 120 125
Thr Ser Thr Phe Phe Pro Leu Val Val Ser Pro Gly Ser Arg Ala Ala
130 135 140
Val Val Phe Ser Ile Cys Leu Asp Lys Ser Met Ser Ser Glu Gly Lys
145 150 155 160
Asp Leu Gln Leu Pro Glu Ser Ile Leu Asn Ile Lys Tyr Gly Ile His
165 170 175
Gly Asp Arg Ala Ala Gly Ala His Arg Pro Val Asp Ala Asp His Ser
180 185 190
Glu Thr Asp Thr
195

(2) INFORMATION FOR SEQ ID NO:105:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 149 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..149
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481641

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:105:

Met Arg Thr Val Ala Leu Lys Leu Glu Phe Gly Val His His Asn Gln
1 5 10 15
Ile Phe Glu Arg Thr Ile Ala Ala His Phe Thr Asp Pro Phe Asp Val
20 25 30
Thr Thr Arg Val Ala Asn Lys Cys Asn Asp Gly Thr Leu Val Leu Gln
35 40 45
Val Met Leu His Ser Leu Val Lys Ala Asn Leu Ile Val Leu Asp Val
50 55 60
Trp Leu Asp Leu Gln Asp Gly Phe Ile His Gly Gln Asn Asp Gly Arg
65 70 75 80
Pro Thr Ser Thr Phe Phe Pro Leu Val Val Ser Pro Gly Ser Arg Ala
85 90 95
Ala Val Val Phe Ser Ile Cys Leu Asp Lys Ser Met Ser Ser Glu Gly
100 105 110
Lys Asp Leu Gln Leu Pro Glu Ser Ile Leu Asn Ile Lys Tyr Gly Ile
115 120 125
His Gly Asp Arg Ala Ala Gly Ala His Arg Pro Val Asp Ala Asp His
130 135 140
Ser Glu Thr Asp Thr
145

(2) INFORMATION FOR SEQ ID NO:106:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 110 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..110
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481642

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:106:

Met Gln Trp Arg Val Glu Arg Leu Lys Asp Leu Lys Glu Ser Glu Ala
1 5 10 15
Val Glu Gln Gln His Asp Glu Val Leu Tyr Glu Val Asn Ala Asn Ser

						20						25						30	
Glu	Asn	Trp	Met	Ile	Ala	Gly	Arg	Lys	Arg	Gly	His	Val	Ser	Leu	Ser				
						35						40						45	
Glu	Glu	Gln	Gly	Ser	Arg	Val	Val	Ile	Ser	Ile	Leu	Cys	Val	Pro	Leu				
						50						55						60	
Val	Ala	Gly	Tyr	Val	Arg	Pro	Pro	Gln	Leu	Gly	Leu	Pro	Asn	Val	Glu				
						65						70						75	
Glu	Ala	Asn	Val	Ser	Ser	Asn	Pro	Ser	Gly	Pro	His	Leu	Val	Cys	Val				
						85						90						95	
Leu	Pro	Pro	Leu	Ser	Ser	Ser	Tyr	Cys	Val	Pro	Val	Lys							
						100						105						110	

(2) INFORMATION FOR SEQ ID NO:107:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1337 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
(B) LOCATION: 1..1337
(D) OTHER INFORMATION: / Ceres Seq. ID 1481647

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:107:

aaaaaaaaataa	aaataaaaaa	tcttcacggt	tcttctctct	ctctctctct	cgagccacca	60
aatctgaatt	agggtttttt	gagaatatct	atcttttgat	ttcaaattct	tcacccactg	120
tgtaatttca	ctcgctcagga	ttcatcagag	gaatcatgat	tacagattcg	atcaccaacg	180
cttctgctac	ttcagctccg	agagattccg	gaaagaagaa	gaggaacaat	aagtcggcta	240
agatgaagca	gaacaagctt	ggtctccgtc	gtgagcaatg	gctttctcaa	gttgcggtga	300
gcaataagga	agttaaagag	gagaggagtg	ttaatcgtag	tcaaaagcct	gatcatgaga	360
gttcagataa	ggtgcgtaga	gaagaggata	acaatggttg	gaataatctt	cttcatcatg	420
agagttttat	ggagtcacct	tcaaatagct	ctgttggtgg	tacatatctg	agcactaact	480
tcagtgggag	aagtagcagg	agttagtagta	gcagcagtg	cttttgctct	ggtaataata	540
cagaagagga	aaatgtagac	gatgatgatg	atgggtgtgt	ggatgattgg	gaagctggtg	600
ctgatgcgtt	agcggctgag	gaagagattg	agaaaaagag	tcgtcctctt	gagtcctgtg	660
aagagcaagt	gagtgttgga	caatcagctt	ctaattgtgt	tgatbcgtcg	attagtgatg	720
catcagatgt	tgtgggtggt	gaagatccaa	agcaggaatg	cttgagagtg	tcatcaagga	780
agcagactag	taatatagct	tggaggctag	atgatagcct	tcgcccacag	gggttaccta	840
atttggcgaa	gcagcttagt	tttccggagt	tagacaagcg	tttttagctct	gtggcgattc	900
cgtcttcatt	tcccatatgc	tacgaagact	tggacttgac	ggattcgaat	ttcctccctt	960
gtccttggtg	atttcggctc	tgtctgttct	gccacaagac	catttgcgat	ggagatgggc	1020
gttgccagg	ctgcaggaaa	ccctatgaac	ggaatatggt	caaggctgag	actagtattc	1080
aagggtggtg	tctaacaatt	cggttggctc	gttcgtctag	catgttttgc	aagttttaaa	1140
aggagaggtg	cggtttttct	aacctagtgt	tcttttgtaa	ctcgagaact	tgagctctgt	1200
tttctatgtc	atcttatggt	ctaagtctga	aacactgtgg	tgatgatgta	gaatgtgatg	1260
tgtgaataca	taaaggttgg	tacagaaaat	gattcaaata	catttagata	gtttcaataa	1320
tgaatgctat	gttctcc					

(2) INFORMATION FOR SEQ ID NO:108:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 327 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
(B) LOCATION: 1..327
(D) OTHER INFORMATION: / Ceres Seq. ID 1481648

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:108:

Met Ile Thr Asp Ser Ile Thr Asn Ala Ser Ala Thr Ser Ala Pro Arg

1 5 10 15
Asp Ser Gly Lys Lys Lys Arg Asn Asn Lys Ser Ala Lys Met Lys Gln
20 25 30
Asn Lys Leu Gly Leu Arg Arg Glu Gln Trp Leu Ser Gln Val Ala Val
35 40 45
Ser Asn Lys Glu Val Lys Glu Glu Arg Ser Val Asn Arg Ser Gln Lys
50 55 60
Pro Asp His Glu Ser Ser Asp Lys Val Arg Arg Glu Glu Asp Asn Asn
65 70 75 80
Gly Gly Asn Asn Leu Leu His His Glu Ser Phe Met Glu Ser Pro Ser
85 90 95
Asn Ser Ser Val Gly Gly Thr Tyr Ser Ser Thr Asn Phe Ser Gly Arg
100 105 110
Ser Ser Arg Ser Ser Ser Ser Ser Gly Phe Cys Ser Gly Asn Ile
115 120 125
Thr Glu Glu Glu Asn Val Asp Asp Asp Asp Gly Cys Val Asp Asp
130 135 140
Trp Glu Ala Val Ala Asp Ala Leu Ala Ala Glu Glu Glu Ile Glu Lys
145 150 155 160
Lys Ser Arg Pro Leu Glu Ser Val Lys Glu Gln Val Ser Val Gly Gln
165 170 175
Ser Ala Ser Asn Val Cys Asp Xaa Ser Ile Ser Asp Ala Ser Asp Val
180 185 190
Val Gly Val Glu Asp Pro Lys Gln Glu Cys Leu Arg Val Ser Ser Arg
195 200 205
Lys Gln Thr Ser Asn Arg Ala Trp Arg Leu Asp Asp Asp Leu Arg Pro
210 215 220
Gln Gly Leu Pro Asn Leu Ala Lys Gln Leu Ser Phe Pro Glu Leu Asp
225 230 235 240
Lys Arg Phe Ser Ser Val Ala Ile Pro Ser Ser Cys Pro Ile Cys Tyr
245 250 255
Glu Asp Leu Asp Leu Thr Asp Ser Asn Phe Leu Pro Cys Pro Cys Gly
260 265 270
Phe Arg Leu Cys Leu Phe Cys His Lys Thr Ile Cys Asp Gly Asp Gly
275 280 285
Arg Cys Pro Gly Cys Arg Lys Pro Tyr Glu Arg Asn Met Val Lys Ala
290 295 300
Glu Thr Ser Ile Gln Gly Gly Gly Leu Thr Ile Arg Leu Ala Arg Ser
305 310 315 320
Ser Ser Met Phe Cys Lys Phe
325

(2) INFORMATION FOR SEQ ID NO:109:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 298 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..298
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481649

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:109:

Met Lys Gln Asn Lys Leu Gly Leu Arg Arg Glu Gln Trp Leu Ser Gln
1 5 10 15
Val Ala Val Ser Asn Lys Glu Val Lys Glu Glu Arg Ser Val Asn Arg
20 25 30
Ser Gln Lys Pro Asp His Glu Ser Ser Asp Lys Val Arg Arg Glu Glu
35 40 45

Asp Asn Asn Gly Gly Asn Asn Leu Leu His His Glu Ser Phe Met Glu
50 55 60
Ser Pro Ser Asn Ser Ser Val Gly Gly Thr Tyr Ser Ser Thr Asn Phe
65 70 75 80
Ser Gly Arg Ser Ser Arg Ser Ser Ser Ser Ser Gly Phe Cys Ser
85 90 95
Gly Asn Ile Thr Glu Glu Glu Asn Val Asp Asp Asp Asp Asp Gly Cys
100 105 110
Val Asp Asp Trp Glu Ala Val Ala Asp Ala Leu Ala Ala Glu Glu Glu
115 120 125
Ile Glu Lys Lys Ser Arg Pro Leu Glu Ser Val Lys Glu Gln Val Ser
130 135 140
Val Gly Gln Ser Ala Ser Asn Val Cys Asp Xaa Ser Ile Ser Asp Ala
145 150 155 160
Ser Asp Val Val Gly Val Glu Asp Pro Lys Gln Glu Cys Leu Arg Val
165 170 175
Ser Ser Arg Lys Gln Thr Ser Asn Arg Ala Trp Arg Leu Asp Asp Asp
180 185 190
Leu Arg Pro Gln Gly Leu Pro Asn Leu Ala Lys Gln Leu Ser Phe Pro
195 200 205
Glu Leu Asp Lys Arg Phe Ser Ser Val Ala Ile Pro Ser Ser Cys Pro
210 215 220
Ile Cys Tyr Glu Asp Leu Asp Leu Thr Asp Ser Asn Phe Leu Pro Cys
225 230 235 240
Pro Cys Gly Phe Arg Leu Cys Leu Phe Cys His Lys Thr Ile Cys Asp
245 250 255
Gly Asp Gly Arg Cys Pro Gly Cys Arg Lys Pro Tyr Glu Arg Asn Met
260 265 270
Val Lys Ala Glu Thr Ser Ile Gln Gly Gly Gly Leu Thr Ile Arg Leu
275 280 285
Ala Arg Ser Ser Ser Met Phe Cys Lys Phe
290 295

(2) INFORMATION FOR SEQ ID NO:110:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 236 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..236
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481650

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:110:

Met Glu Ser Pro Ser Asn Ser Ser Val Gly Gly Thr Tyr Ser Ser Thr
1 5 10 15
Asn Phe Ser Gly Arg Ser Ser Arg Ser Ser Ser Ser Ser Gly Phe
20 25 30
Cys Ser Gly Asn Ile Thr Glu Glu Glu Asn Val Asp Asp Asp Asp
35 40 45
Gly Cys Val Asp Asp Trp Glu Ala Val Ala Asp Ala Leu Ala Ala Glu
50 55 60
Glu Glu Ile Glu Lys Lys Ser Arg Pro Leu Glu Ser Val Lys Glu Gln
65 70 75 80
Val Ser Val Gly Gln Ser Ala Ser Asn Val Cys Asp Xaa Ser Ile Ser
85 90 95
Asp Ala Ser Asp Val Val Gly Val Glu Asp Pro Lys Gln Glu Cys Leu
100 105 110
Arg Val Ser Ser Arg Lys Gln Thr Ser Asn Arg Ala Trp Arg Leu Asp

115	120	125
Asp Asp Leu Arg Pro Gln Gly Leu Pro Asn Leu Ala Lys Gln Leu Ser		
130	135	140
Phe Pro Glu Leu Asp Lys Arg Phe Ser Ser Val Ala Ile Pro Ser Ser		
145	150	155
Cys Pro Ile Cys Tyr Glu Asp Leu Asp Leu Thr Asp Ser Asn Phe Leu		160
	165	170
Pro Cys Pro Cys Gly Phe Arg Leu Cys Leu Phe Cys His Lys Thr Ile		175
	180	185
Cys Asp Gly Asp Gly Arg Cys Pro Gly Cys Arg Lys Pro Tyr Glu Arg		190
	195	200
Asn Met Val Lys Ala Glu Thr Ser Ile Gln Gly Gly Gly Leu Thr Ile		205
	210	215
Arg Leu Ala Arg Ser Ser Met Phe Cys Lys Phe		220
225	230	235

(2) INFORMATION FOR SEQ ID NO:111:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1298 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1298
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481668

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:111:

amatcctmat	cgaaaaacgg	aaattagttt	acaggctgta	atttttcttt	caggctctct	60
ctcttttcgtc	gccgaaccag	ttccagaaag	gctgtagcga	ttcaaaattt	cacaaattaa	120
agtcttcttc	ctctcsaat	cagagattgt	ctccttctta	gtcagatct	gggagcttct	180
tgtgatagat	ttggaagaag	atgactgtga	tcatattct	gactagagtt	gactcgatct	240
gtaagaagta	cgacaagtac	gatgtcgaca	agcagcggga	ggccaatata	tccggcgatg	300
atgcctttgc	tcgtctctat	ggagctttcg	aaacccaaat	cgagaccgct	ctcgagaaag	360
ctgaacttgt	tacgaaggag	aaaaacaggg	ctgctgctgt	tgcaatgaat	gctgagatcc	420
gccggacca	ggcacgattg	tcagaggaag	ttcccaagtt	gcaaagactt	gctgtcaagc	480
gggttaagg	ccttacaacc	gaagagcttg	ctgcgagaaa	tgatttggtg	ctcgctcttc	540
cagccaggat	tgaagccata	cctgatggga	cagcaggtgg	ccctaaaagc	actagtgc	600
ggactccctc	ctcaacaaca	tctcgtcctg	atatcaaat	tgattcagat	gggcgttttg	660
acgatgatta	ctttcaagaa	tcaaatgaat	ctagccaatt	caggcaggag	tatgagatgc	720
ggaaaataaa	acaggaacaa	ggtcttgaca	tgatctccga	agggttagat	gctttgaaga	780
acatggcttc	tgatatgaac	gaggaactgg	atagacaagt	tccactgatg	gatgaaatcg	840
acacaaaggt	ggacagagca	acctccgatc	ttaagaacac	caatgttaga	cttaaagata	900
ccgtgaacca	gctgagatct	agccggaact	tctgtatcga	tattgttttg	ttgtgtattg	960
ttctgggtat	cgctgcatac	ttatacaatg	tactgaagta	atgagatgaa	ccctacgaaa	1020
ggacccatta	gtacttatca	cccagtgcaa	tatccagtgt	gtgcttggtg	cttactcttc	1080
ttctctgata	tttctacgag	agtttcttct	taatgtcaag	aatattcaag	tcttatcttc	1140
ctgcatcgac	ttttctccat	gttgctcgtg	tgcatagatt	tcatctgtca	aaatgtgcgt	1200
caaactaatt	gattgctgtg	tctgcggcag	tgtgctatta	ttttccagcc	aaaatatgat	1260
tttttattta	ttttaaaatc	aagccaaatt	ttaattcc			

(2) INFORMATION FOR SEQ ID NO:112:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 266 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..266

(D) OTHER INFORMATION: / Ceres Seq. ID 1481669

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:112:

```
Met Thr Val Ile Asp Ile Leu Thr Arg Val Asp Ser Ile Cys Lys Lys
1      5      10      15
Tyr Asp Lys Tyr Asp Val Asp Lys Gln Arg Glu Ala Asn Ile Ser Gly
20      25      30
Asp Asp Ala Phe Ala Arg Leu Tyr Gly Ala Phe Glu Thr Gln Ile Glu
35      40      45
Thr Ala Leu Glu Lys Ala Glu Leu Val Thr Lys Glu Lys Asn Arg Ala
50      55      60
Ala Ala Val Ala Met Asn Ala Glu Ile Arg Arg Thr Lys Ala Arg Leu
65      70      75      80
Ser Glu Glu Val Pro Lys Leu Gln Arg Leu Ala Val Lys Arg Val Lys
85      90      95
Gly Leu Thr Thr Glu Glu Leu Ala Ala Arg Asn Asp Leu Val Leu Ala
100     105     110
Leu Pro Ala Arg Ile Glu Ala Ile Pro Asp Gly Thr Ala Gly Gly Pro
115     120     125
Lys Ser Thr Ser Ala Trp Thr Pro Ser Ser Thr Thr Ser Arg Pro Asp
130     135     140
Ile Lys Phe Asp Ser Asp Gly Arg Phe Asp Asp Asp Tyr Phe Gln Glu
145     150     155     160
Ser Asn Glu Ser Ser Gln Phe Arg Gln Glu Tyr Glu Met Arg Lys Ile
165     170     175
Lys Gln Glu Gln Gly Leu Asp Met Ile Ser Glu Gly Leu Asp Ala Leu
180     185     190
Lys Asn Met Ala Ser Asp Met Asn Glu Glu Leu Asp Arg Gln Val Pro
195     200     205
Leu Met Asp Glu Ile Asp Thr Lys Val Asp Arg Ala Thr Ser Asp Leu
210     215     220
Lys Asn Thr Asn Val Arg Leu Lys Asp Thr Val Asn Gln Leu Arg Ser
225     230     235     240
Ser Arg Asn Phe Cys Ile Asp Ile Val Leu Leu Cys Ile Val Leu Gly
245     250     255
Ile Ala Ala Tyr Leu Tyr Asn Val Leu Lys
260     265
```

(2) INFORMATION FOR SEQ ID NO:113:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 198 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..198

(D) OTHER INFORMATION: / Ceres Seq. ID 1481670

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:113:

```
Met Asn Ala Glu Ile Arg Arg Thr Lys Ala Arg Leu Ser Glu Glu Val
1      5      10      15
Pro Lys Leu Gln Arg Leu Ala Val Lys Arg Val Lys Gly Leu Thr Thr
20      25      30
Glu Glu Leu Ala Ala Arg Asn Asp Leu Val Leu Ala Leu Pro Ala Arg
35      40      45
Ile Glu Ala Ile Pro Asp Gly Thr Ala Gly Gly Pro Lys Ser Thr Ser
50      55      60
Ala Trp Thr Pro Ser Ser Thr Thr Ser Arg Pro Asp Ile Lys Phe Asp
65      70      75      80
Ser Asp Gly Arg Phe Asp Asp Asp Tyr Phe Gln Glu Ser Asn Glu Ser
```

```

      85      90      95
Ser Gln Phe Arg Gln Glu Tyr Glu Met Arg Lys Ile Lys Gln Glu Gln
      100      105      110
Gly Leu Asp Met Ile Ser Glu Gly Leu Asp Ala Leu Lys Asn Met Ala
      115      120      125
Ser Asp Met Asn Glu Glu Leu Asp Arg Gln Val Pro Leu Met Asp Glu
      130      135      140
Ile Asp Thr Lys Val Asp Arg Ala Thr Ser Asp Leu Lys Asn Thr Asn
      145      150      155      160
Val Arg Leu Lys Asp Thr Val Asn Gln Leu Arg Ser Ser Arg Asn Phe
      165      170      175
Cys Ile Asp Ile Val Leu Leu Cys Ile Val Leu Gly Ile Ala Ala Tyr
      180      185      190
Leu Tyr Asn Val Leu Lys
      195
```

(2) INFORMATION FOR SEQ ID NO:114:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 770 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..770
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481681

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:114:

```

cytttcgacc tctcctactt actactactc agctgcttct gccttctagg gttctttctc      60
cgttcaccct ccgccgcacg agttgtccag ctccgccgca ttcttctgtc tcccagatca      120
ccggctttta gcaaatccgg ctgctttttca ctctaattcg taaaccactt gtgggatttg      180
agcatctttt acattctcca aaatctctgc tttctagggt tttgtgagtt ttgggtgggat      240
gagtagtggt ttcagtgatc agatcctgat tgataagctc gctaagctca atagcagtca      300
acagtctatc gaaactctgt cacattgggt tatattcaat cggagcaaag cagaattgat      360
cgttacgaca tgggagaaac agtttcacag tacagagatg gatcagaaag tccctctttt      420
gtatttggtt aatgatattc ttcagaacag taagcgtcaa ggtaatgagt ttgtgcaaga      480
gttctggaat gttcttccta aggcctctta agacattggt tctcaaggag atgataatgg      540
caaaagcgct gtcgcacgtg tgatcaagat atgggaagaa agaagagtgt ttggatcacg      600
ttcaaagagt cttaaagatg taatgcttgg agaagatggt cctctgccac ttgatatcag      660
caaaaagcgg gsctcgcgga tccaaatctt caaaacggga gtcaaaatcg tccagaacga      720
aattaacatc aagtgtggtg gtgctgarar gtagcatcag catatcattt
```

(2) INFORMATION FOR SEQ ID NO:115:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 171 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..171
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481682

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:115:

```

Met Ser Ser Val Phe Ser Asp Gln Ile Leu Ile Asp Lys Leu Ala Lys
1      5      10      15
Leu Asn Ser Ser Gln Gln Ser Ile Glu Thr Leu Ser His Trp Cys Ile
      20      25      30
Phe Asn Arg Ser Lys Ala Glu Leu Ile Val Thr Thr Trp Glu Lys Gln
      35      40      45
Phe His Ser Thr Glu Met Asp Gln Lys Val Pro Leu Leu Tyr Leu Ala
```

```

      50              55              60
Asn Asp Ile Leu Gln Asn Ser Lys Arg Gln Gly Asn Glu Phe Val Gln
65              70              75              80
Glu Phe Trp Asn Val Leu Pro Lys Ala Leu Lys Asp Ile Val Ser Gln
      85              90              95
Gly Asp Asp Asn Gly Lys Ser Ala Val Ala Arg Val Ile Lys Ile Trp
      100              105              110
Glu Glu Arg Arg Val Phe Gly Ser Arg Ser Lys Ser Leu Lys Asp Val
      115              120              125
Met Leu Gly Glu Asp Val Pro Leu Pro Leu Asp Ile Ser Lys Lys Arg
      130              135              140
Xaa Ser Arg Ile Gln Ile Phe Lys Thr Gly Val Lys Ile Val Gln Asn
145              150              155              160
Glu Ile Asn Ile Lys Trp Trp Cys Ala Xaa Xaa
      165              170
```

(2) INFORMATION FOR SEQ ID NO:116:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 118 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..118
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481683

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:116:

```

Met Asp Gln Lys Val Pro Leu Leu Tyr Leu Ala Asn Asp Ile Leu Gln
1              5              10              15
Asn Ser Lys Arg Gln Gly Asn Glu Phe Val Gln Glu Phe Trp Asn Val
      20              25              30
Leu Pro Lys Ala Leu Lys Asp Ile Val Ser Gln Gly Asp Asp Asn Gly
      35              40              45
Lys Ser Ala Val Ala Arg Val Ile Lys Ile Trp Glu Glu Arg Arg Val
      50              55              60
Phe Gly Ser Arg Ser Lys Ser Leu Lys Asp Val Met Leu Gly Glu Asp
65              70              75              80
Val Pro Leu Pro Leu Asp Ile Ser Lys Lys Arg Xaa Ser Arg Ile Gln
      85              90              95
Ile Phe Lys Thr Gly Val Lys Ile Val Gln Asn Glu Ile Asn Ile Lys
      100              105              110
Trp Trp Cys Ala Xaa Xaa
      115
```

(2) INFORMATION FOR SEQ ID NO:117:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1004 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1004
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481700

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:117:

```

ctcgcacgc atcgatcctc ccatctgcgc acccgcaagc ctattctccg cacctcctca      60
ggtgaccggg aagatgatgc cgttgagcca aaccgacttc tcgccgtcgc agttcacctc      120
ctcccagaat gccgccgccg actccaccac gccttccaag atgcgcggcg cgtccagcac      180
catgccgctc accgtgaagc aggtcgtcga cgcgcasagt ctggcacggg cgagaagggc      240
```

```
gctccgttca tcgtcaatgg cgtcgagatg gctaacattc gacttgtggg gatgggtcaat 300
gccaaggtgg agcggacgac cgatgtgacc ttcacgctcg acgatggcac cggccgcctc 360
gatttcatca gatgggtgaa tgatgcttca gattcttttg aaactgctgc tattcagaat 420
ggtatgtaca ttgcggtcat tggaagcctc aagggactgc aagagaggaa gcgtgctact 480
gctttctcaa tcaggcctat aaccgatttc aatgaggtta cgctgcattt cattcagtgt 540
gttcggatgc atatagagaa cattgaatta aaggctggca gtcctgcacg aatcagttct 600
tctatgggag tgtcattctc aaatggattc agtgaatcaa gcacaccgac atctttgaaa 660
tccagtcccg caccggtgac cagcgggtca tccgatactg atctgcacac gcaggtcctg 720
aattttttta atgaaccagc gaacctcgag agtgagcatg ggggtgcacgt tgatgaagta 780
ctcaagcggg tcaaactttt gccgaagaag cagatcacgg atgctattga ttacaatatg 840
gactcggggc gtctttactc aacaattgat gaattocact acaaggcaac ttaaccgatt 900
tgaaggccag cctgctggaa atggcagagg actaagtatc acttgtacta aaccaaagtc 960
tgaaatgtc atgttgtgtc atgaaatgca tggttggttt atgg
```

(2) INFORMATION FOR SEQ ID NO:118:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 297 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..297
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481701

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:118:

```
Leu Ala Ser His Arg Ser Ser His Leu Arg Thr Arg Lys Pro Ile Leu
1          5          10          15
Arg Thr Ser Ser Gly Asp Arg Glu Asp Asp Ala Val Glu Pro Asn Arg
20          25          30
Leu Leu Ala Val Ala Val His Leu Leu Pro Glu Cys Arg Arg Arg Leu
35          40          45
His His Ala Phe Gln Asp Ala Arg Arg Val Gln His His Ala Ala His
50          55          60
Arg Glu Ala Gly Arg Arg Arg Ala Xaa Ser Gly Thr Gly Glu Lys Gly
65          70          75          80
Ala Pro Phe Ile Val Asn Gly Val Glu Met Ala Asn Ile Arg Leu Val
85          90          95
Gly Met Val Asn Ala Lys Val Glu Arg Thr Thr Asp Val Thr Phe Thr
100         105         110
Leu Asp Asp Gly Thr Gly Arg Leu Asp Phe Ile Arg Trp Val Asn Asp
115         120         125
Ala Ser Asp Ser Phe Glu Thr Ala Ala Ile Gln Asn Gly Met Tyr Ile
130         135         140
Ala Val Ile Gly Ser Leu Lys Gly Leu Gln Glu Arg Lys Arg Ala Thr
145         150         155         160
Ala Phe Ser Ile Arg Pro Ile Thr Asp Phe Asn Glu Val Thr Leu His
165         170         175
Phe Ile Gln Cys Val Arg Met His Ile Glu Asn Ile Glu Leu Lys Ala
180         185         190
Gly Ser Pro Ala Arg Ile Ser Ser Ser Met Gly Val Ser Phe Ser Asn
195         200         205
Gly Phe Ser Glu Ser Ser Thr Pro Thr Ser Leu Lys Ser Ser Pro Ala
210         215         220
Pro Val Thr Ser Gly Ser Ser Asp Thr Asp Leu His Thr Gln Val Leu
225         230         235         240
Asn Phe Phe Asn Glu Pro Ala Asn Leu Glu Ser Glu His Gly Val His
245         250         255
Val Asp Glu Val Leu Lys Arg Phe Lys Leu Leu Pro Lys Lys Gln Ile
260         265         270
```

Thr Asp Ala Ile Asp Tyr Asn Met Asp Ser Gly Arg Leu Tyr Ser Thr
275 280 285
Ile Asp Glu Phe His Tyr Lys Ala Thr
290 295

(2) INFORMATION FOR SEQ ID NO:119:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 208 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..208

(D) OTHER INFORMATION: / Ceres Seq. ID 1481702

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:119:

Met Ala Asn Ile Arg Leu Val Gly Met Val Asn Ala Lys Val Glu Arg
1 5 10 15
Thr Thr Asp Val Thr Phe Thr Leu Asp Asp Gly Thr Gly Arg Leu Asp
20 25 30
Phe Ile Arg Trp Val Asn Asp Ala Ser Asp Ser Phe Glu Thr Ala Ala
35 40 45
Ile Gln Asn Gly Met Tyr Ile Ala Val Ile Gly Ser Leu Lys Gly Leu
50 55 60
Gln Glu Arg Lys Arg Ala Thr Ala Phe Ser Ile Arg Pro Ile Thr Asp
65 70 75 80
Phe Asn Glu Val Thr Leu His Phe Ile Gln Cys Val Arg Met His Ile
85 90 95
Glu Asn Ile Glu Leu Lys Ala Gly Ser Pro Ala Arg Ile Ser Ser Ser
100 105 110
Met Gly Val Ser Phe Ser Asn Gly Phe Ser Glu Ser Ser Thr Pro Thr
115 120 125
Ser Leu Lys Ser Ser Pro Ala Pro Val Thr Ser Gly Ser Ser Asp Thr
130 135 140
Asp Leu His Thr Gln Val Leu Asn Phe Phe Asn Glu Pro Ala Asn Leu
145 150 155 160
Glu Ser Glu His Gly Val His Val Asp Glu Val Leu Lys Arg Phe Lys
165 170 175
Leu Leu Pro Lys Lys Gln Ile Thr Asp Ala Ile Asp Tyr Asn Met Asp
180 185 190
Ser Gly Arg Leu Tyr Ser Thr Ile Asp Glu Phe His Tyr Lys Ala Thr
195 200 205

(2) INFORMATION FOR SEQ ID NO:120:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 200 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..200

(D) OTHER INFORMATION: / Ceres Seq. ID 1481703

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:120:

Met Val Asn Ala Lys Val Glu Arg Thr Thr Asp Val Thr Phe Thr Leu
1 5 10 15
Asp Asp Gly Thr Gly Arg Leu Asp Phe Ile Arg Trp Val Asn Asp Ala

	20		25		30										
Ser	Asp	Ser	Phe	Glu	Thr	Ala	Ala	Ile	Gln	Asn	Gly	Met	Tyr	Ile	Ala
	35						40					45			
Val	Ile	Gly	Ser	Leu	Lys	Gly	Leu	Gln	Glu	Arg	Lys	Arg	Ala	Thr	Ala
	50					55					60				
Phe	Ser	Ile	Arg	Pro	Ile	Thr	Asp	Phe	Asn	Glu	Val	Thr	Leu	His	Phe
	65				70				75					80	
Ile	Gln	Cys	Val	Arg	Met	His	Ile	Glu	Asn	Ile	Glu	Leu	Lys	Ala	Gly
			85						90					95	
Ser	Pro	Ala	Arg	Ile	Ser	Ser	Ser	Met	Gly	Val	Ser	Phe	Ser	Asn	Gly
		100						105				110			
Phe	Ser	Glu	Ser	Ser	Thr	Pro	Thr	Ser	Leu	Lys	Ser	Ser	Pro	Ala	Pro
	115					120					125				
Val	Thr	Ser	Gly	Ser	Ser	Asp	Thr	Asp	Leu	His	Thr	Gln	Val	Leu	Asn
	130					135					140				
Phe	Phe	Asn	Glu	Pro	Ala	Asn	Leu	Glu	Ser	Glu	His	Gly	Val	His	Val
	145				150				155					160	
Asp	Glu	Val	Leu	Lys	Arg	Phe	Lys	Leu	Leu	Pro	Lys	Lys	Gln	Ile	Thr
			165					170					175		
Asp	Ala	Ile	Asp	Tyr	Asn	Met	Asp	Ser	Gly	Arg	Leu	Tyr	Ser	Thr	Ile
		180				185						190			
Asp	Glu	Phe	His	Tyr	Lys	Ala	Thr								
	195					200									

(2) INFORMATION FOR SEQ ID NO:121:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 500 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..500
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481704

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:121:

atcattactc	cactccacat	tcgacaaaat	atatcttaga	caagttaagt	ttaacgataa	60
tggattcaag	atatgttacc	ctatgcattt	tcttagtact	tgccctacat	ggagatacta	120
ctttggcaga	aacttgcagg	cagtatgttg	aagggcagcc	attttgcttt	aaagcaatgt	180
gcaaggcaaa	ttgttttatg	gagggaaaat	tctctgatgg	ttcttatgta	aagggttaca	240
gatgtgaatc	aggtggattc	cactcgggtg	gtgtttgcct	tttgtgcaa	aattagttat	300
ctaaagacaa	gcggatatat	cttcttatgt	tcctatccat	tatttaggat	tatagtccaa	360
ataattatac	aatagcttag	ttaaatagtt	ttttatttat	agacaaatgt	agcactagtt	420
aactagttgt	gatttttttaa	atttctcagc	tataaatcag	gaaatatattt	ttaacacttc	480
aataatatat	ctttgttcgc					

(2) INFORMATION FOR SEQ ID NO:122:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 49 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..49
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481705

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:122:

Ser	Leu	Leu	His	Thr	Phe	Asp	Lys	Ile	Tyr	Leu	Arg	Gln	Val	Lys	
1			5				10				15				
Phe	Asn	Asp	Asn	Gly	Phe	Lys	Ile	Cys	Tyr	Pro	Met	His	Phe	Leu	Ser

20 25 30
Thr Cys Leu Thr Trp Arg Tyr Tyr Phe Gly Arg Asn Leu Gln Ala Val
 35 40 45
Cys

(2) INFORMATION FOR SEQ ID NO:123:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 97 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..97
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481706

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:123:

His Tyr Ser Thr Pro His Ser Thr Lys Tyr Ile Leu Asp Lys Leu Ser
1 5 10 15
Leu Thr Ile Met Asp Ser Arg Tyr Val Thr Leu Cys Ile Phe Leu Val
 20 25 30
Leu Ala Leu His Gly Asp Thr Thr Leu Ala Glu Thr Cys Arg Gln Tyr
 35 40 45
Val Glu Gly Gln Pro Phe Cys Phe Lys Ala Met Cys Lys Ala Asn Cys
 50 55 60
Phe Met Glu Gly Lys Phe Ser Asp Gly Ser Tyr Val Lys Gly Tyr Arg
65 70 75 80
Cys Glu Ser Gly Gly Phe His Ser Val Cys Val Cys Leu Leu Cys Lys
 85 90 95
Asn

(2) INFORMATION FOR SEQ ID NO:124:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 78 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..78
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481707

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:124:

Met Asp Ser Arg Tyr Val Thr Leu Cys Ile Phe Leu Val Leu Ala Leu
1 5 10 15
His Gly Asp Thr Thr Leu Ala Glu Thr Cys Arg Gln Tyr Val Glu Gly
 20 25 30
Gln Pro Phe Cys Phe Lys Ala Met Cys Lys Ala Asn Cys Phe Met Glu
 35 40 45
Gly Lys Phe Ser Asp Gly Ser Tyr Val Lys Gly Tyr Arg Cys Glu Ser
50 55 60
Gly Gly Phe His Ser Val Cys Val Cys Leu Leu Cys Lys Asn
65 70 75

(2) INFORMATION FOR SEQ ID NO:125:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 916 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..916

(D) OTHER INFORMATION: / Ceres Seq. ID 1481716

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:125:

aatttctgag	caccaaacca	accaagccaa	tcttacgact	gccttgccct	tggcatgtct	60
ctaatttcga	cagcttcgat	actcctcctc	cacgcctgcw	ccctgctcgc	cgccgcaaga	120
gtccccgacc	cggtagaaca	tggcgacgtg	aacacggcga	tgcttaccaa	cggtcgggcc	180
tcggcgacgc	cgtcgtcccc	tgacagcagc	agcaacggca	acttcgagac	gtacttctgc	240
ttcctctgct	cgggcccgcga	cccgtctgct	attcaccact	gccccatcta	ctgggacgag	300
tgccacctca	tctgcgacga	tgacatgtcc	accgccactc	ctactccacc	tgctgttgca	360
gtgtcgctgt	cgtcgtcgct	ccmgccccgt	ccccatgggt	caggtgcagg	gcgatgatga	420
ctgtacgtc	atgaagctct	acatgtccgg	ccgtacgtc	atcgtcgaac	accggccatg	480
caaatacact	gacctggtgt	tcctcacntg	cggcgscggg	gagctggcgg	cggccgaccg	540
gaaagccgtc	acggccactg	cgatccaggg	gacctctctg	cctgccgagc	tatgcggcac	600
gcaggcggtc	aatgctccac	cattagcagg	cgtcgtcgct	ccagcagcag	cagcagcagc	660
tggtggtgct	ggtgcgccac	gacggcgcta	gctgcctagc	tacttatccg	cgaactaagg	720
gttaatttta	gacataaaac	ctgagaggag	gattcaaggg	attaaaatct	ctttcttatt	780
ccaaagaaat	tttagccact	cgaatcctct	ctgattttct	ggctccctaa	ttagccctaa	840
tattaaagga	ccaacagatg	ccaaacttaa	accatcgatc	tctacaagat	aactaatatt	900
ttgatttcca	aacttt					

(2) INFORMATION FOR SEQ ID NO:126:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 120 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..120

(D) OTHER INFORMATION: / Ceres Seq. ID 1481717

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:126:

Met	Ser	Leu	Ile	Ser	Thr	Ala	Ser	Ile	Leu	Leu	Leu	His	Ala	Cys	Xaa
1				5					10					15	
Leu	Leu	Ala	Ala	Ala	Arg	Val	Pro	Asp	Pro	Val	Glu	His	Gly	Asp	Val
				20				25					30		
Asn	Thr	Ala	Met	Leu	Thr	Asn	Gly	Ser	Ala	Ser	Ala	Thr	Pro	Ser	Ser
			35				40					45			
Pro	Asp	Ser	Ser	Ser	Asn	Gly	Asn	Phe	Glu	Thr	Tyr	Phe	Cys	Phe	Leu
			50			55					60				
Cys	Ser	Gly	Arg	Asp	Pro	Leu	Leu	Ile	His	His	Cys	Pro	Ile	Tyr	Trp
65				70					75					80	
Asp	Glu	Cys	His	Leu	Ile	Cys	Asp	Asp	Asp	Met	Ser	Thr	Ala	Thr	Pro
			85					90						95	
Thr	Pro	Pro	Ala	Val	Ala	Val	Ser	Ser	Ser	Ser	Ser	Ser	Xaa	Pro	Arg
			100				105						110		
Pro	His	Gly	Ala	Gly	Ala	Gly	Arg								
			115				120								

(2) INFORMATION FOR SEQ ID NO:127:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 123 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..123

(D) OTHER INFORMATION: / Ceres Seq. ID 1481718

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:127:

```
Met Thr Cys Pro Pro Leu Leu Leu His Leu Leu Leu Gln Cys Arg
1          5          10          15
Arg Arg Arg Arg Pro Xaa Pro Val Pro Met Val Gln Val Gln Gly Asp
          20          25          30
Asp Asp Cys Tyr Val Met Lys Leu Tyr Met Ser Gly Arg Tyr Val Ile
          35          40          45
Val Glu His Arg Pro Cys Lys Tyr Ile Ala Trp Cys Phe Leu Xaa Cys
          50          55          60
Gly Xaa Gly Glu Leu Ala Ala Asp Arg Lys Ala Val Thr Ala Thr
65          70          75          80
Ala Ile Gln Gly Thr Ser Leu Pro Ala Glu Leu Cys Gly Thr Gln Ala
          85          90          95
Val Asn Ala Pro Pro Leu Ala Gly Val Val Val Pro Ala Ala Ala Ala
          100          105          110
Ala Ala Gly Gly Ala Gly Ala His Arg Arg Arg
          115          120
```

(2) INFORMATION FOR SEQ ID NO:128:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 98 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..98

(D) OTHER INFORMATION: / Ceres Seq. ID 1481719

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:128:

```
Met Val Gln Val Gln Gly Asp Asp Asp Cys Tyr Val Met Lys Leu Tyr
1          5          10          15
Met Ser Gly Arg Tyr Val Ile Val Glu His Arg Pro Cys Lys Tyr Ile
          20          25          30
Ala Trp Cys Phe Leu Xaa Cys Gly Xaa Gly Glu Leu Ala Ala Asp
          35          40          45
Arg Lys Ala Val Thr Ala Thr Ala Ile Gln Gly Thr Ser Leu Pro Ala
          50          55          60
Glu Leu Cys Gly Thr Gln Ala Val Asn Ala Pro Pro Leu Ala Gly Val
65          70          75          80
Val Val Pro Ala Ala Ala Ala Ala Ala Gly Gly Ala Gly Ala His Arg
          85          90          95
Arg Arg
```

(2) INFORMATION FOR SEQ ID NO:129:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 553 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..553

(D) OTHER INFORMATION: / Ceres Seq. ID 1481728

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:129:

```
aaatcgcggt cactgctccg aagtccgaac cttcatacac atcgtctcgt ttccgatttc      60
cccaaattca ggccacaggc gctacaggac ccaggcacca ctcggtcggc ggccaccgcg      120
```

```
tcgccccgcc tgctcgattg gggtcgcgtg tgcgatagga agtattgtgt tgtgtttgca 180
acgtgatagc ttgtactggg aacaaagggtc aagatgggcg ccttggacct acaccttgac 240
tttgcttctg ctcaacatgg acaagccaag ttaaaggaat atgccaagag ctctctgttg 300
tctgatggaa actacaatac agacaagatc aatgggttcaa accctgatga ctatgagaaa 360
tttgagaaaag ggataatgca ctatgggtgt ccacattata gaaggagatg ccgcataaga 420
gctccttgct gcaatgaaat ttttgattgc cgacactgcc acaatgaaac taagaattcc 480
attaaaattg ataaaatgaa gaggcgatgaa ctccacgcgc atgaagtgca gcaggttgta 540
tgctcattgt gtg
```

(2) INFORMATION FOR SEQ ID NO:130:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 61 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..61

(D) OTHER INFORMATION: / Ceres Seq. ID 1481729

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:130:

```
Lys Ser Arg Ser Leu Leu Arg Ser Pro Asn Leu His Thr His Arg Leu
1          5          10          15
Val Ser Asp Phe Pro Lys Phe Arg Pro Gln Ala Leu Gln Asp Pro Gly
20          25          30
Thr Thr Arg Ser Ala Ala Thr Ala Ser Pro Arg Leu Leu Asp Trp Gly
35          40          45
Arg Val Cys Asp Arg Lys Tyr Cys Val Val Phe Ala Thr
50          55          60
```

(2) INFORMATION FOR SEQ ID NO:131:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 90 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..90

(D) OTHER INFORMATION: / Ceres Seq. ID 1481730

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:131:

```
Asn Arg Gly His Cys Ser Glu Val Arg Thr Phe Ile His Ile Val Ser
1          5          10          15
Phe Pro Ile Ser Pro Asn Ser Gly His Arg Arg Tyr Arg Thr Gln Ala
20          25          30
Pro Leu Gly Arg Arg Pro Pro Arg Arg Pro Ala Cys Ser Ile Gly Val
35          40          45
Ala Cys Ala Ile Gly Ser Ile Val Leu Cys Leu Gln Arg Asp Ser Leu
50          55          60
Tyr Trp Glu Gln Arg Ser Arg Trp Ala Pro Trp Thr Tyr Thr Leu Thr
65          70          75          80
Leu Leu Leu Leu Asn Met Asp Lys Pro Ser
85          90
```

(2) INFORMATION FOR SEQ ID NO:132:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 113 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..113

(D) OTHER INFORMATION: / Ceres Seq. ID 1481731

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:132:

```
Met Gly Ala Leu Asp Leu His Leu Asp Phe Ala Ser Ala Gln His Gly
1           5           10           15
Gln Ala Lys Leu Lys Glu Tyr Ala Lys Ser Ser Leu Leu Ser Asp Gly
20           25           30
Asn Tyr Asn Thr Asp Lys Ile Asn Gly Ser Asn Pro Asp Asp Tyr Glu
35           40           45
Lys Phe Glu Lys Gly Ile Met His Tyr Gly Cys Pro His Tyr Arg Arg
50           55           60
Arg Cys Arg Ile Arg Ala Pro Cys Cys Asn Glu Ile Phe Asp Cys Arg
65           70           75           80
His Cys His Asn Glu Thr Lys Asn Ser Ile Lys Ile Asp Lys Met Lys
85           90           95
Arg His Glu Leu Pro Arg His Glu Val Gln Gln Val Val Cys Ser Leu
100          105          110
Cys
```

(2) INFORMATION FOR SEQ ID NO:133:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 709 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..709

(D) OTHER INFORMATION: / Ceres Seq. ID 1481732

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:133:

```
attttcttcg tgccctggcg ttcccagggtg ccatgcgagt agatcggcaa ctactccatc 60
ctcctctccc tacctggcca tcgtgcagca gccgtgctga gcctctgctg ccctctctct 120
ggccactcgc gtgagcccct gctggtcggc tgtccggaca cgacggctat ggccgagccc 180
agcgcgaagt catcctccag gactgagccc ctgtgcagcc actaccgccg gcaggcctcg 240
gtgtccgggt gtgcttcgac aagtcacgcg cctccgacaa acctcgcgcc cgtcgtgttc 300
ttgcctctac gtcaatcgac acgcatagaa tgcgatccat ctccaagctt cgagggatcg 360
aagatcaagc gttggcgacc atgtgatcaa gctctctgag ttctatgagg ctgaagatcc 420
tgagcatctg tttggtgaag attgcctttg gtgcaatcta tgctcaggta aagagggcgt 480
cgaggcggat ctccaggagt tccaggacgt cgacgggttc gaggattagg ctacgcacct 540
ccccagtcg gctgcctgtg gtgggttggt tacgttggct acgtttcgat tctgtgtact 600
ttgatttata ttatgtaaat ggttctagtt tgtaatatta ttacttactc tttattgtaa 660
ttcgaagcat tgtgctatga tgagtcattt atgtaatcgc cgtgtacgc
```

(2) INFORMATION FOR SEQ ID NO:134:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 43 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..43

(D) OTHER INFORMATION: / Ceres Seq. ID 1481733

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:134:

```
Phe Leu Arg Ala Trp Ala Phe Pro Gly Ala Met Arg Val Asp Arg Gln
1           5           10           15
```

Leu Leu His Pro Pro Leu Pro Thr Trp Pro Ser Cys Ser Ser Arg Ala
20 25 30
Glu Pro Leu Leu Pro Ser Leu Trp Pro Leu Ala
35 40

(2) INFORMATION FOR SEQ ID NO:135:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 76 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..76
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481734

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:135:

Met Ala Glu Pro Ser Ala Lys Ser Ser Ser Arg Thr Glu Pro Leu Cys
1 5 10 15
Ser His Tyr Arg Arg Gln Ala Ser Val Ser Gly Cys Ala Ser Thr Ser
20 25 30
His Ala Pro Pro Thr Asn Leu Ala Pro Val Val Phe Leu Pro Leu Arg
35 40 45
Gln Ser Thr Arg Ile Glu Cys Asp Pro Ser Pro Ser Phe Glu Gly Ser
50 55 60
Lys Ile Lys Arg Trp Arg Pro Cys Asp Gln Ala Leu
65 70 75

(2) INFORMATION FOR SEQ ID NO:136:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 76 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..76
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481735

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:136:

Met Arg Leu Lys Ile Leu Ser Ile Cys Leu Val Lys Ile Ala Phe Gly
1 5 10 15
Ala Ile Tyr Ala Gln Val Lys Arg Ala Ser Arg Arg Ile Ser Arg Ser
20 25 30
Ser Arg Thr Ser Thr Gly Ser Arg Ile Arg Leu Ala Thr Ser Pro Ser
35 40 45
Gln Leu Pro Val Val Gly Cys Leu Arg Trp Leu Arg Phe Asp Ser Val
50 55 60
Tyr Phe Asp Leu Tyr Tyr Val Asn Gly Ser Ser Leu
65 70 75

(2) INFORMATION FOR SEQ ID NO:137:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 951 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..951
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481740

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:137:

attacacaaa	tgtgcgcgc	catgttctcc	aatctcttcg	ccaagtttga	ctacggacga	60
tcgtctccac	cgaagacgcc	acacgatgac	ggccgccgta	gccacatgtc	tgatctttcc	120
ctagaaagac	agcctcgacg	gtcgtccgtc	tccgtccgca	tgaggcgcc	cgtggatgat	180
gacgacgtca	ctgcggcgcc	cgtagccgag	gtgatgagca	cggaccatgg	cggccacgag	240
gagtcgtctc	caccgaagac	gccacacgat	gacggctgcc	gtagccacat	gtctgatctt	300
tccctagaaa	gacagcctcg	acagtcgtcc	gtctcgggcc	gcatggaggc	gcccgtggat	360
gacgacgacg	tcactkcggc	gcccgtagcc	gaggtgatga	gcatggacca	tggcggccac	420
gaggagtgcg	cgacgggtccc	gtgcctcgcg	ttcgcgtccg	agcacgggta	cagcatcttc	480
tccctagcct	acatgcgcgd	tgttcatcga	cggcgscac	ggkttcamag	tcaccgccga	540
cccagwggga	gcgaaagcga	aaccgcgkt	acgtgattct	tgccaaycgg	ctaacacmcc	600
catktggacg	tctggccgtc	gtgtttgacg	tcggttctc	cgaccttdgg	aggccagagc	660
scwtggggcg	gctaaagcta	aacmccggcg	aggttgasc	aatktggggc	cagccgcact	720
ggatcatgcc	trgggataga	tccgatcgtc	gtcaaggata	twtcaactag	tacagtttat	780
tgtaggtagt	tmcattagtt	tacatactct	ggctgtcagg	cmctatttct	acgtaaagtt	840
ttttttggca	ttrgggaaat	atattmcgga	tctataagat	atthttgrgtt	ttaaaagcta	900
ctgataaaatc	tacatgtacg	ttgcaatgcg	aaataaactg	tgtctatgtt	t	

(2) INFORMATION FOR SEQ ID NO:138:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 191 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..191
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481741

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:138:

Ile	Thr	Gln	Met	Cys	Ala	Ala	Met	Phe	Ser	Asn	Leu	Phe	Ala	Lys	Phe	
1				5					10					15		
Asp	Tyr	Gly	Arg	Ser	Ser	Pro	Pro	Lys	Thr	Pro	His	Asp	Asp	Gly	Arg	
			20					25					30			
Arg	Ser	His	Met	Ser	Asp	Leu	Ser	Leu	Glu	Arg	Gln	Pro	Arg	Arg	Ser	
		35					40					45				
Ser	Val	Ser	Val	Arg	Met	Glu	Ala	Pro	Val	Asp	Asp	Asp	Asp	Val	Thr	
	50					55				60						
Ala	Ala	Pro	Val	Ala	Glu	Val	Met	Ser	Thr	Asp	His	Gly	Gly	His	Glu	
65					70					75				80		
Glu	Ser	Ser	Pro	Pro	Lys	Thr	Pro	His	Asp	Asp	Gly	Cys	Arg	Ser	His	
			85						90				95			
Met	Ser	Asp	Leu	Ser	Leu	Glu	Arg	Gln	Pro	Arg	Gln	Ser	Ser	Val	Ser	
			100					105					110			
Val	Arg	Met	Glu	Ala	Pro	Val	Asp	Asp	Asp	Asp	Val	Thr	Xaa	Ala	Pro	
	115						120					125				
Val	Ala	Glu	Val	Met	Ser	Met	Asp	His	Gly	Gly	His	Glu	Glu	Ser	Pro	
	130					135					140					
Thr	Val	Pro	Cys	Leu	Ala	Phe	Ala	Ser	Glu	His	Gly	Tyr	Ser	Ile	Phe	
145				150						155					160	
Ser	Leu	Ala	Tyr	Met	Arg	Xaa	Val	His	Arg	Arg	Arg	Xaa	Arg	Xaa	Xaa	
			165					170					175			
Ser	His	Arg	Arg	Pro	Xaa	Gly	Ser	Glu	Ser	Glu	Thr	Ala	Xaa	Thr		
	180						185						190			

(2) INFORMATION FOR SEQ ID NO:139:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 188 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..188
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481742
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:139:

```
Met Cys Ala Ala Met Phe Ser Asn Leu Phe Ala Lys Phe Asp Tyr Gly
1          5          10          15
Arg Ser Ser Pro Pro Lys Thr Pro His Asp Asp Gly Arg Arg Ser His
          20          25          30
Met Ser Asp Leu Ser Leu Glu Arg Gln Pro Arg Arg Ser Ser Val Ser
          35          40          45
Val Arg Met Glu Ala Pro Val Asp Asp Asp Asp Val Thr Ala Ala Pro
          50          55          60
Val Ala Glu Val Met Ser Thr Asp His Gly Gly His Glu Glu Ser Ser
          65          70          75          80
Pro Pro Lys Thr Pro His Asp Asp Gly Cys Arg Ser His Met Ser Asp
          85          90          95
Leu Ser Leu Glu Arg Gln Pro Arg Gln Ser Ser Val Ser Val Arg Met
          100         105         110
Glu Ala Pro Val Asp Asp Asp Asp Val Thr Xaa Ala Pro Val Ala Glu
          115         120         125
Val Met Ser Met Asp His Gly Gly His Glu Glu Ser Pro Thr Val Pro
          130         135         140
Cys Leu Ala Phe Ala Ser Glu His Gly Tyr Ser Ile Phe Ser Leu Ala
          145         150         155         160
Tyr Met Arg Xaa Val His Arg Arg Arg Xaa Arg Xaa Xaa Ser His Arg
          165         170         175
Arg Pro Xaa Gly Ser Glu Ser Glu Thr Ala Xaa Thr
          180         185
```

(2) INFORMATION FOR SEQ ID NO:140:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 184 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

- (ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..184
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481743
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:140:

```
Met Phe Ser Asn Leu Phe Ala Lys Phe Asp Tyr Gly Arg Ser Ser Pro
1          5          10          15
Pro Lys Thr Pro His Asp Asp Gly Arg Arg Ser His Met Ser Asp Leu
          20          25          30
Ser Leu Glu Arg Gln Pro Arg Arg Ser Ser Val Ser Val Arg Met Glu
          35          40          45
Ala Pro Val Asp Asp Asp Asp Val Thr Ala Ala Pro Val Ala Glu Val
          50          55          60
Met Ser Thr Asp His Gly Gly His Glu Glu Ser Ser Pro Pro Lys Thr
          65          70          75          80
Pro His Asp Asp Gly Cys Arg Ser His Met Ser Asp Leu Ser Leu Glu
          85          90          95
Arg Gln Pro Arg Gln Ser Ser Val Ser Val Arg Met Glu Ala Pro Val
          100         105         110
Asp Asp Asp Asp Val Thr Xaa Ala Pro Val Ala Glu Val Met Ser Met
          115         120         125
Asp His Gly Gly His Glu Glu Ser Pro Thr Val Pro Cys Leu Ala Phe
```

130 135 140
Ala Ser Glu His Gly Tyr Ser Ile Phe Ser Leu Ala Tyr Met Arg Xaa
145 150 155 160
Val His Arg Arg Arg Xaa Arg Xaa Xaa Ser His Arg Arg Pro Xaa Gly
165 170 175
Ser Glu Ser Glu Thr Ala Xaa Thr
180

(2) INFORMATION FOR SEQ ID NO:141:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 432 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..432
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481744

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:141:

agttctaaac cctaaacctg acgcccgcacat ggccgcgcgc gttcgccaca tcgtgcgcgc	60
ccgcctctcc acggccgcgc ccatactctg accggtcccc actccggcct ccatactcaa	120
cccgtcctcg ccgagcactc ccctcacctc gcgacataag acccgactcg ccatactcct	180
cctcaagtct tccccgcgc ctccccccga ccagatcctc tccatttgcc gcgcgcgcga	240
ctgaccccg agacacacat cgaccgcac gcgctgtcgc tagccgcac aaagctctcc	300
tccgctcccg acaccctccg tgacctcgcc tccacmgccc tcaccccgcg cmamgcaccc	360
cacgcmatcg cgctcttcgg ccaggcacam ctctccccg acgssatctc cactttccag	420
tcctccccct cc	

(2) INFORMATION FOR SEQ ID NO:142:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 80 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..80
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481745

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:142:

Val Leu Asn Pro Lys Pro Asp Ala Ala Met Ala Ala Val Arg His	
1 5 10 15	
Ile Val Arg Arg Arg Leu Ser Thr Ala Ala Ala Ile Thr Ala Pro Val	
20 25 30	
Pro Thr Pro Ala Ser Ile Leu Asn Pro Ser Ser Pro Ser Thr Pro Leu	
35 40 45	
Thr Ser Arg His Lys Thr Arg Leu Ala Ile Ser Leu Leu Lys Ser Ser	
50 55 60	
Pro Pro Pro Pro Asp Gln Ile Leu Ser Ile Cys Arg Ala Ala His	
65 70 75 80	

(2) INFORMATION FOR SEQ ID NO:143:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 71 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide
(B) LOCATION: 1..71
(D) OTHER INFORMATION: / Ceres Seq. ID 1481746

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:143:

```
Met Ala Ala Ala Val Arg His Ile Val Arg Arg Arg Leu Ser Thr Ala
1          5          10          15
Ala Ala Ile Thr Ala Pro Val Pro Thr Pro Ala Ser Ile Leu Asn Pro
20          25          30
Ser Ser Pro Ser Thr Pro Leu Thr Ser Arg His Lys Thr Arg Leu Ala
35          40          45
Ile Ser Leu Leu Lys Ser Ser Pro Pro Pro Pro Pro Asp Gln Ile Leu
50          55          60
Ser Ile Cys Arg Ala Ala His
65          70
```

(2) INFORMATION FOR SEQ ID NO:144:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 557 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -
(B) LOCATION: 1..557
(D) OTHER INFORMATION: / Ceres Seq. ID 1481747

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:144:

```
agactccggc cacagccgag acgagactag cagcagccgc ttgctcagat cggcagcttc      60
ggcggcgggcg gagatggcga ttcgggtactg gccgatggcc ggagcagccg ttgggttccg    120
cctcgctcctg gttctcttcg gcggggatct ccaccttgcc tctcgccctg aggtctccac    180
ccccctcacc tcccttcgcc gcctggcgga aggctactgg ctgaagcaag cgtccgtgtc     240
accgtactcc ggttctatgt atcacggttc cccattgtct ctgtctgttc ttggtccatt     300
aactagtagc aggctgacg gacatcatgc tcatatttac tgcagtttga tttttgtggc     360
tgtagatttt ctagcagcca tgctcatccg agcgactggg catgaactcg aaatggcacg     420
gaacagaagt ttgaagtcac ttgacctcac aaaggcagtw aaggatacag ttaatgtaag     480
cgctggagat gttgcttctc tcatatatatt gtggaaccct tgggcaatag tcacttgtgt     540
gggatcatgt acatcac
```

(2) INFORMATION FOR SEQ ID NO:145:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 185 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide
(B) LOCATION: 1..185
(D) OTHER INFORMATION: / Ceres Seq. ID 1481748

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:145:

```
Asp Ser Gly His Ser Arg Asp Glu Thr Ser Ser Arg Leu Leu Arg
1          5          10          15
Ser Ala Ala Ser Ala Ala Ala Glu Met Ala Ile Arg Tyr Trp Pro Met
20          25          30
Ala Gly Ala Ala Val Gly Phe Arg Leu Val Leu Val Leu Phe Gly Gly
35          40          45
Asp Leu His Leu Ala Ser Arg Pro Glu Val Ser Thr Pro Leu Thr Ser
50          55          60
Leu Arg Arg Leu Ala Glu Gly Tyr Trp Leu Lys Gln Ala Ser Val Ser
65          70          75          80
Pro Tyr Ser Gly Ser Met Tyr His Gly Ser Pro Leu Leu Leu Ser Val
```

			85				90				95				
Leu	Gly	Pro	Leu	Thr	Ser	Ser	Arg	Pro	Asp	Gly	His	His	Ala	His	Ile
			100					105					110		
Tyr	Cys	Ser	Leu	Ile	Phe	Val	Ala	Val	Asp	Phe	Leu	Ala	Ala	Met	Leu
		115					120					125			
Ile	Arg	Ala	Thr	Gly	His	Glu	Leu	Glu	Met	Ala	Arg	Asn	Arg	Ser	Leu
		130				135					140				
Lys	Ser	Leu	Asp	Leu	Thr	Lys	Ala	Xaa	Lys	Asp	Thr	Val	Asn	Val	Ser
145					150					155				160	
Ala	Gly	Asp	Val	Ala	Ser	Leu	Ile	Tyr	Leu	Trp	Asn	Pro	Trp	Ala	Ile
			165					170						175	
Val	Thr	Cys	Val	Gly	Ser	Cys	Thr	Ser							
			180				185								

(2) INFORMATION FOR SEQ ID NO:146:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 161 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..161

(D) OTHER INFORMATION: / Ceres Seq. ID 1481749

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:146:

Met	Ala	Ile	Arg	Tyr	Trp	Pro	Met	Ala	Gly	Ala	Ala	Val	Gly	Phe	Arg
1				5					10					15	
Leu	Val	Leu	Val	Leu	Phe	Gly	Gly	Asp	Leu	His	Leu	Ala	Ser	Arg	Pro
			20					25					30		
Glu	Val	Ser	Thr	Pro	Leu	Thr	Ser	Leu	Arg	Arg	Leu	Ala	Glu	Gly	Tyr
		35				40						45			
Trp	Leu	Lys	Gln	Ala	Ser	Val	Ser	Pro	Tyr	Ser	Gly	Ser	Met	Tyr	His
	50					55					60				
Gly	Ser	Pro	Leu	Leu	Leu	Ser	Val	Leu	Gly	Pro	Leu	Thr	Ser	Ser	Arg
65					70				75					80	
Pro	Asp	Gly	His	His	Ala	His	Ile	Tyr	Cys	Ser	Leu	Ile	Phe	Val	Ala
			85					90						95	
Val	Asp	Phe	Leu	Ala	Ala	Met	Leu	Ile	Arg	Ala	Thr	Gly	His	Glu	Leu
			100					105					110		
Glu	Met	Ala	Arg	Asn	Arg	Ser	Leu	Lys	Ser	Leu	Asp	Leu	Thr	Lys	Ala
		115					120					125			
Xaa	Lys	Asp	Thr	Val	Asn	Val	Ser	Ala	Gly	Asp	Val	Ala	Ser	Leu	Ile
	130					135					140				
Tyr	Leu	Trp	Asn	Pro	Trp	Ala	Ile	Val	Thr	Cys	Val	Gly	Ser	Cys	Thr
145					150					155				160	
Ser															

(2) INFORMATION FOR SEQ ID NO:147:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 154 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..154

(D) OTHER INFORMATION: / Ceres Seq. ID 1481750

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:147:

Met Ala Gly Ala Ala Val Gly Phe Arg Leu Val Leu Val Leu Phe Gly
1 5 10 15
Gly Asp Leu His Leu Ala Ser Arg Pro Glu Val Ser Thr Pro Leu Thr
20 25 30
Ser Leu Arg Arg Leu Ala Glu Gly Tyr Trp Leu Lys Gln Ala Ser Val
35 40 45
Ser Pro Tyr Ser Gly Ser Met Tyr His Gly Ser Pro Leu Leu Leu Ser
50 55 60
Val Leu Gly Pro Leu Thr Ser Ser Arg Pro Asp Gly His His Ala His
65 70 75 80
Ile Tyr Cys Ser Leu Ile Phe Val Ala Val Asp Phe Leu Ala Ala Met
85 90 95
Leu Ile Arg Ala Thr Gly His Glu Leu Glu Met Ala Arg Asn Arg Ser
100 105 110
Leu Lys Ser Leu Asp Leu Thr Lys Ala Xaa Lys Asp Thr Val Asn Val
115 120 125
Ser Ala Gly Asp Val Ala Ser Leu Ile Tyr Leu Trp Asn Pro Trp Ala
130 135 140
Ile Val Thr Cys Val Gly Ser Cys Thr Ser
145 150

(2) INFORMATION FOR SEQ ID NO:148:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 380 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..380
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481755

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:148:

acaagcaagt ggccaccttt gagtggatgt tggaagaaat agcagccaca agcaagtagt	60
cacctgtgtc atcttattcc gatacctagc cctcccatct ccaawkcctc gsttcctcct	120
cccttctcta gttctctgat cctcagcact tagcatcaag cttagsacac cggcgagatg	180
gcctccamct ccamcttctt gtccamcctc gccagcaggt ccgcggcagc cgatagcctg	240
ygcamgccgt gccgtccttc gccaagatcg tcaggttctt gcccgcgcar gcgcagatca	300
gccgcavggg cmgcgcggcg gtgctgcccc cgccgarggc ggcgggtgtcg ggcacgagaa	360
ggcgccgctcg agcaagcacg	

(2) INFORMATION FOR SEQ ID NO:149:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 67 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..67
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481756

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:149:

Met Ala Ser Xaa Ser Xaa Phe Leu Ser Xaa Leu Ala Ser Arg Ser Ala
1 5 10 15
Ala Ala Asp Ser Leu Xaa Xaa Pro Cys Arg Pro Ser Pro Arg Ser Ser
20 25 30
Gly Ser Cys Pro Arg Xaa Arg Arg Ser Ala Ala Xaa Xaa Ala Arg Arg
35 40 45
Cys Cys Pro Arg Arg Xaa Arg Arg Cys Arg Ala Arg Glu Gly Ala Val
50 55 60

Glu Gln Ala

65

(2) INFORMATION FOR SEQ ID NO:150:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 482 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..482
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481764

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:150:

ctgttcttcc	acctgctggc	tgcwcttgcc	tcccctgcgc	cccaaaccga	ccgcctcgc	60
cgtccccgca	gccgcagcct	gctctcggt	cccgcgcgcg	tctaccgcgt	cctgcggctg	120
cggtgttgcg	tcacctcggt	ttcgctttaa	cttcacaaat	cctcgccgtc	ctggtgctcc	180
gccgccccctc	cctttgtact	cgcgctggag	ctgcagatcc	accgcgacct	ggcgaccaat	240
tcctcctccc	gctgaagaat	tggcgacctt	ggcctccgcm	cccgcggcgc	gaggagtcaa	300
ctgtggtagc	aaccaccgcg	gaggctgcaa	gcttcggtaa	gggaggaaaag	ttgacttggt	360
ggaagccggt	ccagggccgc	gatgacgtcg	acagccgccg	ggcgctcgctg	tcggcggcga	420
agagcgagtc	ctacctgcgg	gccgacaaga	tcgacctcga	gagcctggac	atccagctgg	480
ag						

(2) INFORMATION FOR SEQ ID NO:151:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 112 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..112
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481765

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:151:

Cys	Ser	Ser	Thr	Cys	Trp	Leu	Xaa	Leu	Pro	Pro	Leu	Arg	Pro	Lys	Pro
1				5					10					15	
Thr	Arg	Leu	Ala	Val	Pro	Ala	Ala	Ala	Cys	Ser	Arg	Leu	Pro	Pro	
			20					25				30			
Pro	Ser	Thr	Ala	Ser	Cys	Gly	Cys	Gly	Val	Ala	Ser	Pro	Arg	Val	Arg
			35				40					45			
Leu	Asn	Phe	His	Asn	Pro	Arg	Arg	Pro	Gly	Ala	Pro	Pro	Pro	Leu	Pro
			50				55					60			
Leu	Tyr	Ser	Arg	Trp	Ser	Cys	Arg	Ser	Thr	Ala	Thr	Trp	Arg	Pro	Ile
65						70				75				80	
Pro	Pro	Pro	Ala	Glu	Glu	Leu	Ala	Thr	Leu	Ala	Ser	Xaa	Pro	Ala	Ala
						85				90				95	
Arg	Gly	Val	Asn	Cys	Gly	Ser	Asn	His	Arg	Gly	Gly	Cys	Lys	Leu	Arg
			100					105						110	

(2) INFORMATION FOR SEQ ID NO:152:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 160 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide
(B) LOCATION: 1..160
(D) OTHER INFORMATION: / Ceres Seq. ID 1481766

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:152:

Val	Leu	Pro	Pro	Ala	Gly	Cys	Xaa	Cys	Leu	Pro	Cys	Ala	Pro	Asn	Pro
1				5					10					15	
Pro	Ala	Ser	Pro	Ser	Pro	Gln	Pro	Gln	Pro	Ala	Leu	Gly	Ser	Arg	Arg
			20					25					30		
Arg	Leu	Pro	Arg	Pro	Ala	Ala	Ala	Val	Leu	Arg	His	Leu	Gly	Phe	Ala
			35				40					45			
Leu	Thr	Ser	Thr	Ile	Leu	Ala	Val	Leu	Val	Leu	Arg	Arg	Pro	Ser	Leu
	50					55				60					
Cys	Thr	Arg	Ala	Gly	Ala	Ala	Asp	Pro	Pro	Arg	Pro	Gly	Asp	Gln	Phe
65					70				75					80	
Leu	Leu	Pro	Leu	Lys	Asn	Trp	Arg	Pro	Trp	Pro	Pro	Xaa	Pro	Arg	Arg
				85					90					95	
Glu	Glu	Ser	Thr	Val	Val	Ala	Thr	Thr	Ala	Glu	Ala	Ala	Ser	Phe	Gly
			100					105					110		
Lys	Gly	Gly	Lys	Leu	Thr	Cys	Trp	Lys	Pro	Val	Gln	Gly	Arg	Asp	Asp
			115				120					125			
Val	Asp	Ser	Arg	Arg	Ala	Ser	Ser	Ser	Ala	Ala	Lys	Ser	Glu	Ser	Tyr
	130					135					140				
Leu	Arg	Ala	Asp	Lys	Ile	Asp	Leu	Glu	Ser	Leu	Asp	Ile	Gln	Leu	Glu
145					150					155					160

(2) INFORMATION FOR SEQ ID NO:153:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 376 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -
(B) LOCATION: 1..376
(D) OTHER INFORMATION: / Ceres Seq. ID 1481770

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:153:

ggactcacga	agcagcacac	tctgcaactct	cggcaacaac	tgacggcccg	aggaagaagg	60
cgcagacgac	aagcagaagc	ttgtgccatc	gatcaatggc	ggcgggtgaca	aagatctacg	120
tcgtgtacta	ctcgacgtac	gghcacgtgg	cgargctggc	ggaggagatc	aagaagggcg	180
ccgactccgt	ggacggcgct	gaggcaacca	tctggcargw	agcggaracg	ctgccggavg	240
argcgctggc	gaagatgcrc	gcaccggcga	ggagcgagga	gcaccgggtg	atctcgggca	300
arcagctggt	ggacgcrzac	ggcatcctgt	tcggcttccc	rgcrcggttc	ggcatgatgg	360
crgcgcatga	gaaggc					

(2) INFORMATION FOR SEQ ID NO:154:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 124 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide
(B) LOCATION: 1..124
(D) OTHER INFORMATION: / Ceres Seq. ID 1481771

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:154:

Thr	His	Glu	Ala	Ala	His	Ser	Ala	Leu	Ser	Ala	Thr	Thr	Asp	Gly	Arg
1				5				10						15	

Arg Lys Lys Ala Gln Thr Thr Ser Arg Ser Leu Cys His Arg Ser Met
20 25 30
Ala Ala Val Thr Lys Ile Tyr Val Val Tyr Tyr Ser Thr Tyr Xaa His
35 40 45
Val Ala Xaa Leu Ala Glu Glu Ile Lys Lys Gly Ala Asp Ser Val Asp
50 55 60
Gly Val Glu Ala Thr Ile Trp Xaa Xaa Ala Xaa Thr Leu Pro Xaa Xaa
65 70 75 80
Ala Leu Ala Lys Met Xaa Ala Pro Ala Arg Ser Glu Glu His Pro Val
85 90 95
Ile Ser Gly Xaa Gln Leu Val Asp Xaa Asp Gly Ile Leu Phe Gly Phe
100 105 110
Xaa Xaa Arg Phe Gly Met Met Xaa Ala Gln Met Lys
115 120

(2) INFORMATION FOR SEQ ID NO:155:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 93 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..93

(D) OTHER INFORMATION: / Ceres Seq. ID 1481772

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:155:

Met Ala Ala Val Thr Lys Ile Tyr Val Val Tyr Tyr Ser Thr Tyr Xaa
1 5 10 15
His Val Ala Xaa Leu Ala Glu Glu Ile Lys Lys Gly Ala Asp Ser Val
20 25 30
Asp Gly Val Glu Ala Thr Ile Trp Xaa Xaa Ala Xaa Thr Leu Pro Xaa
35 40 45
Xaa Ala Leu Ala Lys Met Xaa Ala Pro Ala Arg Ser Glu Glu His Pro
50 55 60
Val Ile Ser Gly Xaa Gln Leu Val Asp Xaa Asp Gly Ile Leu Phe Gly
65 70 75 80
Phe Xaa Xaa Arg Phe Gly Met Met Xaa Ala Gln Met Lys
85 90

(2) INFORMATION FOR SEQ ID NO:156:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 448 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..448

(D) OTHER INFORMATION: / Ceres Seq. ID 1481775

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:156:

attgagtata ggtttgctct cctacacttt tttgagaaag acattgaagg atgacatagt	60
tgttccaatg cttgatttta agatccaaga tggggacatt gtaccgttgg tgtatggttc	120
acagggtgat tgggatagta gtctgaagat agtacttgat tgggtcccctt tttcttcgaa	180
ggaagaactt ctgcagcagt ttcaggatgt tggtagtcat ggaactaaag tggtagtgta	240
caatttatgg atgaatgatg atggcctttt ggaacttgac tttgaggatg atgatgagga	300
catattactt agagatcaag gtagcgcaag tsvggggggt ctcaaagagt cagaaagaaa	360
ttgttaagca acacatatcc cacaggctca gakttttcat tgcgagctta tacctccatc	420
ctttacctca ggaagtttga taatttcc	

(2) INFORMATION FOR SEQ ID NO:157:

(i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 121 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..121
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481776
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:157:
Leu Ser Ile Gly Leu Leu Ser Tyr Thr Phe Leu Arg Lys Thr Leu Lys
1 5 10 15
Asp Asp Ile Val Val Pro Met Leu Asp Phe Lys Ile Gln Asp Gly Asp
 20 25 30
Ile Val Pro Leu Val Tyr Gly Ser Gln Gly Asp Trp Asp Ser Ser Leu
 35 40 45
Lys Ile Val Leu Asp Trp Ser Pro Phe Ser Ser Lys Glu Glu Leu Leu
 50 55 60
Gln Gln Phe Gln Asp Val Gly Ser His Gly Thr Lys Val Val Val Tyr
65 70 75 80
Asn Leu Trp Met Asn Asp Asp Gly Leu Leu Glu Leu Asp Phe Glu Asp
 85 90 95
Asp Asp Glu Asp Ile Leu Leu Arg Asp Gln Gly Ser Ala Ser Xaa Gly
 100 105 110
Val Leu Lys Glu Ser Glu Arg Asn Cys
 115 120

(2) INFORMATION FOR SEQ ID NO:158:

(i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 99 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..99
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481777
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:158:
Met Leu Asp Phe Lys Ile Gln Asp Gly Asp Ile Val Pro Leu Val Tyr
1 5 10 15
Gly Ser Gln Gly Asp Trp Asp Ser Ser Leu Lys Ile Val Leu Asp Trp
 20 25 30
Ser Pro Phe Ser Ser Lys Glu Glu Leu Leu Gln Gln Phe Gln Asp Val
 35 40 45
Gly Ser His Gly Thr Lys Val Val Val Tyr Asn Leu Trp Met Asn Asp
50 55 60
Asp Gly Leu Leu Glu Leu Asp Phe Glu Asp Asp Asp Glu Asp Ile Leu
65 70 75 80
Leu Arg Asp Gln Gly Ser Ala Ser Xaa Gly Val Leu Lys Glu Ser Glu
 85 90 95
Arg Asn Cys

(2) INFORMATION FOR SEQ ID NO:159:

(i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 61 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..61
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481778

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:159:

```
Met Met Met Ala Phe Trp Asn Leu Thr Leu Arg Met Met Met Arg Thr
1           5           10           15
Tyr Tyr Leu Glu Ile Lys Val Ala Gln Xaa Xaa Gly Phe Ser Lys Ser
          20          25          30
Gln Lys Glu Ile Val Lys Gln His Ile Ser His Arg Leu Arg Xaa Phe
        35         40         45
Ile Ala Ser Leu Tyr Leu His Pro Leu Pro Gln Glu Val
       50        55        60
```

(2) INFORMATION FOR SEQ ID NO:160:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 657 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

- (ix) FEATURE:
 (A) NAME/KEY: -
 (B) LOCATION: 1..657
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481779

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:160:

```
attatgggatt agcgcttgca taacagacca ccagggaatt ccgctctcac ctcatggggg      60
tcctggtggg cgaataagca ccaagccctc cgcttggatc grccatcgtg cgcgtccctt      120
gctcgcgtcg agagatttcc aacgccgagc tagcgcacag ggaagggaga gaaaagtgag      180
tgccaccgcg gggacgaggg aaactcgact caccacctat cgctccttgg agttgtgtcc      240
tggggattct ccgggcttgg aggaggaagg ttgaagcagc tcttccgggg attcgtgttc      300
aacgttgttg gaattcctcg ccaccaggaa ctaccctgcc gtggaccgcg ggtccgcgtg      360
gccaagactg tcctcgctgc tagccttgac gagcaagcca cacatgatcg agtgcttcag      420
tagtgggaga ctgctgagca ggagctcatt gaagccccc acaggagaag aacatgtggt      480
tcaagaggag ccacacaatg aggagtcca cttgatctag gtgtcgtttc ccagttgact      540
ttatggcgcc aaggatggac atttgttcgt tttatattat tttttgtaa gacttccgct      600
atgtaataag tactctgatt atattgtgac atttatctct atacactctg ttattgt
```

(2) INFORMATION FOR SEQ ID NO:161:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 90 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

- (ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..90
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481780

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:161:

```
Leu Trp Ile Ser Ala Cys Ile Thr Asp His Gln Gly Ile Pro Leu Ser
1           5           10           15
Pro His Trp Gly Pro Gly Gly Arg Ile Ser Thr Lys Pro Ser Ala Trp
        20         25         30
Ile Xaa His Arg Ala Arg Pro Leu Leu Ala Ser Arg Asp Phe Gln Arg
        35         40         45
Arg Ala Ser Ala Gln Gly Arg Glu Arg Lys Val Ser Ala Thr Ala Gly
       50        55        60
Thr Arg Glu Thr Arg Leu Thr Thr Tyr Arg Ser Leu Glu Leu Cys Pro
      65          70          75          80
```

Gly Asp Ser Pro Gly Leu Glu Glu Gly Gly
85 90

(2) INFORMATION FOR SEQ ID NO:162:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 413 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..413
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481789

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:162:

ctgctcgctc tctccctctc gtcgctttct ttcctgggtcg cgcgcgcctt catcagggtct	60
cctccgcctt agccgggtgaa gagcgaccag gcccaataaa taatcaccat ggctcatcaa	120
aagcgtgaag gcagctacgc cgatgatgat agtacatcca agcgcatcaa aggcaccgac	180
actgcttctg aaacggggga cagtgtagag tctagtgttt cacagcaaat ggatgctgaa	240
gctaggagga cctgccaaaa ggaaagcgaa caccatcgga caaatgcgtt tcagatgggg	300
aatgcgctgc aaactctaag gttttggggg aagcagaaga kggatttgac tgttgctcag	360
gctgatgcgg cggacgacaa gggttgcagg cacactatgg aggacgcctg gggk	

(2) INFORMATION FOR SEQ ID NO:163:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 101 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..101
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481790

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:163:

Met Ala His Gln Lys Arg Glu Gly Ser Tyr Ala Asp Asp Asp Ser Thr	
1 5 10 15	
Ser Lys Arg Ile Lys Gly Thr Asp Thr Ala Ser Glu Thr Gly Asp Ser	
20 25 30	
Val Glu Ser Ser Val Ser Gln Gln Met Asp Ala Glu Ala Arg Arg Thr	
35 40 45	
Cys Gln Lys Glu Ser Glu His His Arg Thr Asn Ala Phe Gln Met Gly	
50 55 60	
Asn Ala Leu Gln Thr Leu Arg Phe Trp Gly Lys Gln Lys Xaa Val Leu	
65 70 75 80	
Thr Val Val Glu Ala Asp Ala Ala Asp Asp Lys Gly Cys Arg His Thr	
85 90 95	
Met Glu Asp Ala Trp	
100	

(2) INFORMATION FOR SEQ ID NO:164:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 61 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..61
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481791

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:164:

Met Asp Ala Glu Ala Arg Arg Thr Cys Gln Lys Glu Ser Glu His His
1 5 10 15
Arg Thr Asn Ala Phe Gln Met Gly Asn Ala Leu Gln Thr Leu Arg Phe
20 25 30
Trp Gly Lys Gln Lys Xaa Val Leu Thr Val Val Glu Ala Asp Ala Ala
35 40 45
Asp Asp Lys Gly Cys Arg His Thr Met Glu Asp Ala Trp
50 55 60

(2) INFORMATION FOR SEQ ID NO:165:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 460 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..460
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481792

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:165:

atcaaggaac agtgcaaact agaagaaacc tccgtcatca gcgacctctc cccaacggcg	60
ccgacgatgg cgcaacagca gacgcagcta accactggct cggggcatcct ggatgccgtc	120
ccgctcttcg tcgtcatcct cctcgcgcc cagctcctgg ccctcgtgtt ctggatgtac	180
aagctggcct ccgagaagca accaccccg aggaagacac agtgacggcg ccgatctacg	240
ccatcggcga gtccttcgct agcctcttta tcggttccat tttcatgtga accagtacc	300
tccagaacat tcaggccgtc aattattcag agatatccat atagtctttc aatttgttt	360
atttatactt attgcatttt gggtattgtt tgataacaac ttagcgatat tctatgaatc	420
actatccgtt tgggtgataa ataaatgttc ctagttag	

(2) INFORMATION FOR SEQ ID NO:166:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 74 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..74
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481793

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:166:

Ile Lys Glu Gln Cys Lys Leu Glu Glu Thr Ser Val Ile Ser Asp Leu
1 5 10 15
Ser Pro Thr Ala Pro Thr Met Ala Gln Gln Gln Thr Gln Leu Thr Thr
20 25 30
Gly Ser Gly Ile Leu Asp Ala Val Pro Leu Phe Val Val Ile Leu Leu
35 40 45
Ala Ala His Val Leu Ala Leu Val Phe Trp Met Tyr Lys Leu Ala Ser
50 55 60
Glu Lys Gln Pro Pro Arg Lys Thr Gln
65 70

(2) INFORMATION FOR SEQ ID NO:167:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 95 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide

(B) LOCATION: 1..95

(D) OTHER INFORMATION: / Ceres Seq. ID 1481794

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:167:

```
Gln Gly Thr Val Gln Thr Arg Arg Asn Leu Arg His Gln Arg Pro Leu
1           5           10           15
Pro Asn Gly Ala Asp Asp Gly Ala Thr Ala Asp Ala Ala Asn His Trp
          20           25           30
Leu Gly His Pro Gly Cys Arg Pro Ala Leu Arg Arg His Pro Pro Arg
          35           40           45
Gly Pro Arg Pro Gly Pro Arg Val Leu Asp Val Gln Ala Gly Phe Arg
          50           55           60
Glu Ala Thr Thr Pro Glu Glu Asp Thr Val Thr Ala Pro Ile Tyr Ala
65           70           75           80
Ile Gly Glu Ser Phe Ala Ser Leu Phe Ile Gly Ser Ile Phe Met
          85           90           95
```

(2) INFORMATION FOR SEQ ID NO:168:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 52 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..52

(D) OTHER INFORMATION: / Ceres Seq. ID 1481795

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:168:

```
Met Ala Gln Gln Gln Thr Gln Leu Thr Thr Gly Ser Gly Ile Leu Asp
1           5           10           15
Ala Val Pro Leu Phe Val Val Ile Leu Leu Ala Ala His Val Leu Ala
          20           25           30
Leu Val Phe Trp Met Tyr Lys Leu Ala Ser Glu Lys Gln Pro Pro Arg
          35           40           45
Arg Lys Thr Gln
50
```

(2) INFORMATION FOR SEQ ID NO:169:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 761 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..761

(D) OTHER INFORMATION: / Ceres Seq. ID 1481796

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:169:

```
aggacaacga gacgtctgam ggcgtgargc tgagacccaaa ggagaacaga nacggarang 60
agcdcr gcg cccgggcgcc ggagcctgga gggtagggga agagaagaga agaggcgatg 120
gcgtcaatcg ggtcctccaa catcggattc cagctgctga agaagtctgg ttggaaggag 180
ggcactggcc ttrgagcgca ggagcaggga aggttggaac ctgtagagac tcgtgttaag 240
aataacaagc gtggtatagg ttctaaagaa ccaaaaccac aacctaaggt tgaggatgac 300
attgaaacac atcctcaaaa gcccaagcag gaaatgcaat caaagaaaag ggcaaaatta 360
gctgcaaaga ggataagaaa actgcaagaa gaggagaagc gcttgaaaga gaaggaattc 420
gragatggct tttttcaggg aatttttggc ctgataatgt ggtaaggcag aaacttcaac 480
acttgacaat gtagctgctg acttttgsct gatatagttg atgtataggc ttgcaaaacg 540
cttggcctac aaaatgttac cccattcatt ccgtggatga tttcacacat gatttgtggc 600
tagggttggc acaaagctgt tcatccatgt agtggacaaa tctagtgtag aattgcccac 660
gctatgtgat cttgtaattt ttatacatTA taaatcttgt ttttttttgt tcatataact 720
```

gtgctcaata ttttgccttg ccaatgcaaa tatttaaattc t

(2) INFORMATION FOR SEQ ID NO:170:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 78 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..78

(D) OTHER INFORMATION: / Ceres Seq. ID 1481797

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:170:

Asp	Asn	Glu	Thr	Ser	Xaa	Gly	Val	Xaa	Leu	Arg	Pro	Lys	Glu	Asn	Arg
1			5				10							15	
Xaa	Gly	Xaa	Glu	Xaa	Xaa	Gly	Pro	Gly	Ala	Gly	Ala	Trp	Arg	Val	Gly
			20				25						30		
Glu	Glu	Lys	Arg	Arg	Gly	Asp	Gly	Val	Asn	Arg	Val	Leu	Gln	His	Arg
			35				40						45		
Ile	Pro	Ala	Ala	Glu	Glu	Val	Trp	Leu	Glu	Gly	Gly	His	Trp	Pro	Xaa
			50				55					60			
Ser	Ala	Gly	Ala	Gly	Lys	Val	Gly	Thr	Cys	Arg	Asp	Ser	Cys		
65					70					75					

(2) INFORMATION FOR SEQ ID NO:171:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 115 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..115

(D) OTHER INFORMATION: / Ceres Seq. ID 1481798

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:171:

Met	Ala	Ser	Ile	Gly	Ser	Ser	Asn	Ile	Gly	Phe	Gln	Leu	Leu	Lys	Lys
1			5						10					15	
Ser	Gly	Trp	Lys	Glu	Gly	Thr	Gly	Leu	Xaa	Ala	Gln	Glu	Gln	Gly	Arg
			20				25						30		
Leu	Glu	Pro	Val	Glu	Thr	Arg	Val	Lys	Asn	Asn	Lys	Arg	Gly	Ile	Gly
			35				40						45		
Ser	Lys	Glu	Pro	Lys	Pro	Gln	Pro	Lys	Val	Glu	Asp	Asp	Ile	Glu	Thr
			50				55				60				
His	Pro	Gln	Lys	Pro	Lys	Gln	Glu	Met	Gln	Ser	Lys	Lys	Arg	Ala	Lys
65					70					75				80	
Leu	Ala	Ala	Lys	Arg	Ile	Arg	Lys	Leu	Gln	Glu	Glu	Glu	Lys	Arg	Leu
			85						90					95	
Lys	Glu	Lys	Glu	Phe	Xaa	Asp	Gly	Phe	Phe	Gln	Gly	Ile	Phe	Gly	Leu
			100					105						110	
Ile	Met	Trp													
			115												

(2) INFORMATION FOR SEQ ID NO:172:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 712 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -
(B) LOCATION: 1..712
(D) OTHER INFORMATION: / Ceres Seq. ID 1481799

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:172:

gtttatcttg	gacaaaggaa	ctaagtgaat	ggactatggt	tacttgagta	gattgaagag	60
tggttatggt	tctaaaacaa	atatagtgtg	ctaatttgca	agatgctacc	attatgacat	120
tgtctgtcac	tgaaatctgc	ttgtcacgcc	agaaatagac	atgctttatt	tttgttttct	180
atccttgtca	atttttccgg	caattgaaat	tgttactgtg	tcaattctta	cagtttgcac	240
agtttgttga	atgtactttt	acttttaccg	tagtacaatg	ctaattgtaga	atactgtaac	300
cagtttgcac	ttgcagggtt	tcttctcaga	catcaatgct	tatgagggtg	agcttggtac	360
tgatgaagag	aagcactgct	tctgccgtga	gtcagacttg	ttagctgtag	ttgaatgaat	420
tttaccatga	aaatttcgga	cttacctggg	aatgctccaa	cggcatgaac	ttatcttgcc	480
cccactttgt	tgtatgccat	ttgaacgttc	gttaattcag	cttctagatg	attgttagtt	540
accgttgatt	tttggtcgcc	ataaactgga	attatgttaa	tgccccattt	ctttacagag	600
gctcttgctc	acgtttggaa	tacgaaactg	tgtgaaaccg	aacttgaaat	gtttacattg	660
cccacttgat	gtttctgctt	ctgaacattc	tttaggcaac	atcctaatta	tt	

(2) INFORMATION FOR SEQ ID NO:173:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 794 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -
(B) LOCATION: 1..794
(D) OTHER INFORMATION: / Ceres Seq. ID 1481800

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:173:

atgcatcttc	ctagtcctag	ccgcagctcc	tcactttcct	gcgcgcgctc	cttctctgtc	60
tcccatggcc	gagcgtcctc	cctgctcctc	tctcctttat	ctccctctcc	atgccggcgc	120
ttccctagct	cgttttccca	tctccctgct	ccttcccatg	gtgcgcaggg	cgtoctcccc	180
tgcactccct	ccctagctcc	ccaccgctcg	ccttctctct	cttccctagtc	gcccgtmtcc	240
agctcgggtc	gcactgcgcg	tccttgcccc	tcgctgtttt	tgtggccagc	cgagctcgcc	300
cagccccctg	cctctccacc	tcataacgc	cctcagccat	ggatgtcgaa	tcccctctct	360
ttagtgcctt	tctttgcagc	ccctgcgtcg	ccgtgcatgt	aagggtgttt	gtctaaatgc	420
tcaagaggag	tgtcgtgtcg	tggacagccc	ttttggcgtc	gtcgggtgtt	tgatgttttg	480
cgcaccccg	ctacgacacc	gtcgaccctc	agtgatattt	cgttcttgct	ttgtcgtttt	540
atcgatcgac	gtctatttgc	taatgtgaag	tgtgtgtatg	tgccatgtgt	tgttgaggag	600
cgacatctgt	ggaatctggg	tgaagaagaa	acagagcacg	tccgacgctt	actagctgct	660
ggtgaaagga	ttgaatcggc	tatcatggtc	gtttagtgtc	gatcgagtca	accttagtcg	720
tggtaaagta	ccattatttc	tgctatttag	ccgatgtatg	agttagatgg	ataaaatagt	780
tacgatgatt	ttcc					

(2) INFORMATION FOR SEQ ID NO:174:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 75 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide
(B) LOCATION: 1..75
(D) OTHER INFORMATION: / Ceres Seq. ID 1481801

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:174:

Met	His	Leu	Pro	Ser	Pro	Ser	Arg	Ser	Ser	Ser	Ser	Ser	Cys	Ala	Arg
1				5				10					15		
Ser	Phe	Ser	Val	Ser	His	Gly	Arg	Ala	Ser	Ser	Leu	Leu	Leu	Ser	Pro
			20					25					30		
Leu	Ser	Pro	Ser	Pro	Cys	Arg	Arg	Phe	Pro	Ser	Ser	Phe	Ser	His	Leu

35 40 45
Pro Ala Pro Ser His Gly Ala Gln Gly Val Leu Pro Ser Thr Pro Ser
50 55 60
Leu Ala Pro His Arg Ser Pro Ser Ser Pro Ser
65 70 75

(2) INFORMATION FOR SEQ ID NO:175:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 75 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..75

(D) OTHER INFORMATION: / Ceres Seq. ID 1481802

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:175:

Met His Leu Pro Ser Pro Ser Arg Ser Ser Ser Ser Ser Cys Ala Arg
1 5 10 15
Ser Phe Ser Val Ser His Gly Arg Ala Ser Ser Leu Leu Leu Ser Pro
20 25 30
Leu Ser Pro Ser Pro Cys Arg Arg Phe Pro Ser Ser Phe Ser His Leu
35 40 45
Pro Ala Pro Ser His Gly Ala Gln Gly Val Leu Pro Ser Thr Pro Ser
50 55 60
Leu Ala Pro His Arg Ser Pro Ser Ser Pro Ser
65 70 75

(2) INFORMATION FOR SEQ ID NO:176:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 64 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..64

(D) OTHER INFORMATION: / Ceres Seq. ID 1481803

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:176:

Cys Ile Phe Leu Val Leu Ala Ala Ala Pro His Leu Pro Ala Arg Ala
1 5 10 15
Pro Ser Leu Ser Pro Met Ala Glu Arg Pro Pro Cys Ser Ser Leu Leu
20 25 30
Tyr Leu Pro Leu His Ala Gly Ala Ser Leu Ala Arg Phe Pro Ile Ser
35 40 45
Leu Leu Leu Pro Met Val Arg Arg Ala Ser Ser Pro Arg Leu Pro Pro
50 55 60

(2) INFORMATION FOR SEQ ID NO:177:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 239 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..239

(D) OTHER INFORMATION: / Ceres Seq. ID 1481808

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:177:

atataa	acaa	gcctat	gtta	gcctg	ctacg	cactg	tggtc	gtgcat	gttt	tcattt	gacc	60
ttgttt	ccat	gatgct	tcga	cgtgtt	acgc	gtgctc	gtgc	ttcttg	ctgt	gtcatc	actg	120
gtccac	attt	tcttgt	ggaa	agtggt	ccctt	tgtacg	agct	tatgaa	acca	gtgtgc	acaa	180
gcgacg	gacg	gatttg	tacc	atccag	naac	gnatag	tgan	tccgtt	tttac	taactc	ctg	

(2) INFORMATION FOR SEQ ID NO:178:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 76 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..76

(D) OTHER INFORMATION: / Ceres Seq. ID 1481809

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:178:

Ile	Asn	Lys	Pro	Met	Leu	Ala	Cys	Tyr	Ala	Leu	Cys	Ser	Cys	Met	Phe	
1				5					10					15		
Ser	Phe	Asp	Leu	Val	Ser	Met	Met	Leu	Arg	Arg	Val	Thr	Arg	Ala	Arg	
			20					25					30			
Ala	Ser	Cys	Cys	Val	Ile	Thr	Gly	Pro	His	Phe	Leu	Val	Glu	Ser	Val	
		35					40						45			
Pro	Leu	Tyr	Glu	Leu	Met	Lys	Pro	Val	Cys	Thr	Ser	Asp	Gly	Arg	Ile	
		50				55					60					
Cys	Thr	Ile	Gln	Xaa	Xaa	Ile	Val	Xaa	Pro	Phe	Tyr					
65				70					75							

(2) INFORMATION FOR SEQ ID NO:179:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 72 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..72

(D) OTHER INFORMATION: / Ceres Seq. ID 1481810

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:179:

Met	Leu	Ala	Cys	Tyr	Ala	Leu	Cys	Ser	Cys	Met	Phe	Ser	Phe	Asp	Leu	
1				5					10					15		
Val	Ser	Met	Met	Leu	Arg	Arg	Val	Thr	Arg	Ala	Arg	Ala	Ser	Cys	Cys	
			20					25					30			
Val	Ile	Thr	Gly	Pro	His	Phe	Leu	Val	Glu	Ser	Val	Pro	Leu	Tyr	Glu	
		35					40						45			
Leu	Met	Lys	Pro	Val	Cys	Thr	Ser	Asp	Gly	Arg	Ile	Cys	Thr	Ile	Gln	
		50				55					60					
Xaa	Xaa	Ile	Val	Xaa	Pro	Phe	Tyr									
65				70												

(2) INFORMATION FOR SEQ ID NO:180:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 62 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide

(B) LOCATION: 1..62

(D) OTHER INFORMATION: / Ceres Seq. ID 1481811

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:180:

Met	Phe	Ser	Phe	Asp	Leu	Val	Ser	Met	Met	Leu	Arg	Arg	Val	Thr	Arg
1				5					10					15	
Ala	Arg	Ala	Ser	Cys	Cys	Val	Ile	Thr	Gly	Pro	His	Phe	Leu	Val	Glu
			20					25					30		
Ser	Val	Pro	Leu	Tyr	Glu	Leu	Met	Lys	Pro	Val	Cys	Thr	Ser	Asp	Gly
			35				40					45			
Arg	Ile	Cys	Thr	Ile	Gln	Xaa	Xaa	Ile	Val	Xaa	Pro	Phe	Tyr		
	50				55						60				

(2) INFORMATION FOR SEQ ID NO:181:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 433 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..433

(D) OTHER INFORMATION: / Ceres Seq. ID 1481815

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:181:

gatggaatca	atttcctcga	tcatttttagc	tcagcaata	tgaatcatga	tccatcaatt	60
gctgcagaaa	gtaagcaaaa	caatgaagat	gaacctttaa	gggaaatgaa	gaataaaaag	120
aagaaatgga	agcaaggtag	tagtagcatt	gaaccaaattg	acattctaga	atcttttccc	180
tcagagaaaag	ctagcttaac	tggtcatttt	ggtaccagca	aagctattgt	gccatctggt	240
gcaaaagaaa	gcatgaacat	agaaaatgag	aatgtgaatg	acggcaagga	gaagaagaga	300
aaggggcaaag	ctaatatgga	agtacctact	gctgaaaagg	acaattctaa	ttgtgataat	360
caaggaattg	atattagtac	ccaagaatca	cttaktkctt	ttgtacaaaa	tgaaggrtg	420
ggtcaggaaa	atg					

(2) INFORMATION FOR SEQ ID NO:182:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 144 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..144

(D) OTHER INFORMATION: / Ceres Seq. ID 1481816

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:182:

Asp	Gly	Ile	Asn	Phe	Leu	Asp	His	Phe	Ser	Cys	Ser	Asn	Met	Asn	His
1				5					10					15	
Asp	Pro	Ser	Ile	Ala	Ala	Glu	Ser	Lys	Gln	Asn	Asn	Glu	Asp	Glu	Pro
			20					25					30		
Leu	Arg	Glu	Met	Lys	Asn	Lys	Lys	Lys	Trp	Lys	Gln	Gly	Thr	Ser	
			35				40					45			
Ser	Ile	Glu	Pro	Asn	Asp	Ile	Leu	Glu	Ser	Phe	Pro	Ser	Glu	Lys	Ala
			50			55				60					
Ser	Leu	Thr	Gly	His	Phe	Gly	Thr	Ser	Lys	Ala	Ile	Val	Pro	Ser	Val
65				70					75					80	
Ala	Lys	Glu	Ser	Met	Asn	Ile	Glu	Asn	Glu	Asn	Val	Asn	Asp	Gly	Lys
			85						90					95	
Glu	Lys	Lys	Arg	Lys	Gly	Lys	Ala	Asn	Met	Glu	Val	Pro	Thr	Ala	Glu
			100				105						110		
Lys	Asp	Asn	Ser	Asn	Cys	Asp	Asn	Gln	Gly	Ile	Asp	Ile	Ser	Thr	Gln
			115				120						125		

(2) INFORMATION FOR SEQ ID NO:183:

- ```

(i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 131 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..131
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481817

```

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:183:

[illegible]

- (2) INFORMATION FOR SEQ ID NO:184:

- ```
(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 109 amino acids
    (B) TYPE: amino acid
    (C) STRANDEDNESS:
    (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
    (A) NAME/KEY: peptide
    (B) LOCATION: 1..109
    (D) OTHER INFORMATION: / Ceres Seq. ID 1481818
```

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:184:

Met	Lys	Asn	Lys	Lys	Lys	Lys	Trp	Lys	Gln	Gly	Thr	Ser	Ser	Ile	Glu
1				5					10					15	
Pro	Asn	Asp	Ile	Leu	Glu	Ser	Phe	Pro	Ser	Glu	Lys	Ala	Ser	Leu	Thr
			20					25					30		
Gly	His	Phe	Gly	Thr	Ser	Lys	Ala	Ile	Val	Pro	Ser	Val	Ala	Lys	Glu
		35					40					45			
Ser	Met	Asn	Ile	Glu	Asn	Glu	Asn	Val	Asn	Asp	Gly	Lys	Glu	Lys	Lys
	50					55					60				
Arg	Lys	Gly	Lys	Ala	Asn	Met	Glu	Val	Pro	Thr	Ala	Glu	Lys	Asp	Asn
65					70					75				80	
Ser	Asn	Cys	Asp	Asn	Gln	Gly	Ile	Asp	Ile	Ser	Thr	Gln	Glu	Ser	Leu
			85						90					95	
Xaa	Xaa	Phe	Val	Gln	Asn	Glu	Arg	Xaa	Gly	Gln	Glu	Asn			

- 100 105
- (2) INFORMATION FOR SEQ ID NO:185:
- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 495 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- (ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..495

(D) OTHER INFORMATION: / Ceres Seq. ID 1481819

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:185:

anrgggaggg	ctgcggttgg	ggaagagaga	gatagaagag	aatcagggta	atgcagatgg	60
gatcgatgga	attcgtagcg	ccggcggttg	aggagctttt	gccggaactt	tccctcgagg	120
agcagccacg	gttgccagaac	caatcccgcg	agcgtgaccg	catccggaag	cgacgtaaca	180
agcactctcc	tctctcccgt	ccgtcggttg	tctcggtaca	gtacgtgatg	gatatgggat	240
cgatgggaat	ggatttcgtg	gcgcgcgcgt	tggaggagct	gctgccggat	ctttcccgcg	300
aggagcagct	acggttgcaa	aacaaatccc	gcgggcgtga	ccgcatctcg	aagccacgta	360
acaagcacgc	tctctgtccc	cgctcgtcgc	cgttctcgga	atgggacggc	aacatcttca	420
aaattcccca	ggctctccac	gctctcgscg	actacaatgc	caggcaccct	ggtggcgagt	480
tcgatgttgt	gaagc					

- (2) INFORMATION FOR SEQ ID NO:186:

- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 148 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: peptide

- (ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..148

(D) OTHER INFORMATION: / Ceres Seq. ID 1481820

- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:186:

Met	Gln	Met	Gly	Ser	Met	Glu	Phe	Val	Ala	Pro	Ala	Leu	Glu	Glu	Leu	
1				5					10					15		
Leu	Pro	Glu	Leu	Ser	Leu	Glu	Glu	Gln	Pro	Arg	Leu	Gln	Asn	Gln	Ser	
				20				25					30			
Arg	Glu	Arg	Asp	Arg	Ile	Arg	Lys	Arg	Arg	Asn	Lys	His	Ser	Pro	Pro	
				35				40				45				
Pro	Arg	Pro	Ser	Leu	Ile	Ser	Val	Gln	Tyr	Val	Met	Asp	Met	Gly	Ser	
				50				55			60					
Met	Gly	Met	Asp	Phe	Val	Ala	Pro	Ala	Leu	Glu	Glu	Leu	Leu	Pro	Asp	
65				70				75						80		
Leu	Ser	Arg	Glu	Glu	Gln	Leu	Arg	Leu	Gln	Asn	Lys	Ser	Arg	Gly	Arg	
				85				90						95		
Asp	Arg	Ile	Ser	Lys	Pro	Arg	Asn	Lys	His	Ala	Pro	Arg	Pro	Arg	Pro	
				100				105					110			
Ser	Pro	Phe	Ser	Glu	Trp	Asp	Gly	Asn	Ile	Phe	Lys	Ile	Pro	Gln	Val	
				115				120				125				
Leu	His	Ala	Leu	Xaa	His	Tyr	Asn	Ala	Arg	His	Pro	Gly	Gly	Glu	Phe	
				130				135				140				
Asp	Val	Val	Lys													

- 145
- (2) INFORMATION FOR SEQ ID NO:187:

- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 146 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:

- (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..146
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481821

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:187:

```
Met Gly Ser Met Glu Phe Val Ala Pro Ala Leu Glu Glu Leu Leu Pro
1          5          10          15
Glu Leu Ser Leu Glu Glu Gln Pro Arg Leu Gln Asn Gln Ser Arg Glu
20          25          30
Arg Asp Arg Ile Arg Lys Arg Arg Asn Lys His Ser Pro Pro Pro Arg
35          40          45
Pro Ser Leu Ile Ser Val Gln Tyr Val Met Asp Met Gly Ser Met Gly
50          55          60
Met Asp Phe Val Ala Pro Ala Leu Glu Glu Leu Leu Pro Asp Leu Ser
65          70          75          80
Arg Glu Glu Gln Leu Arg Leu Gln Asn Lys Ser Arg Gly Arg Asp Arg
85          90          95
Ile Ser Lys Pro Arg Asn Lys His Ala Pro Arg Pro Arg Pro Ser Pro
100          105          110
Phe Ser Glu Trp Asp Gly Asn Ile Phe Lys Ile Pro Gln Val Leu His
115          120          125
Ala Leu Xaa His Tyr Asn Ala Arg His Pro Gly Gly Glu Phe Asp Val
130          135          140
Val Lys
145
```

(2) INFORMATION FOR SEQ ID NO:188:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 143 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..143
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481822

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:188:

```
Met Glu Phe Val Ala Pro Ala Leu Glu Leu Leu Pro Glu Leu Ser
1          5          10          15
Leu Glu Glu Gln Pro Arg Leu Gln Asn Gln Ser Arg Glu Arg Asp Arg
20          25          30
Ile Arg Lys Arg Arg Asn Lys His Ser Pro Pro Pro Arg Pro Ser Leu
35          40          45
Ile Ser Val Gln Tyr Val Met Asp Met Gly Ser Met Gly Met Asp Phe
50          55          60
Val Ala Pro Ala Leu Glu Leu Leu Pro Asp Leu Ser Arg Glu Glu
65          70          75          80
Gln Leu Arg Leu Gln Asn Lys Ser Arg Gly Arg Asp Arg Ile Ser Lys
85          90          95
Pro Arg Asn Lys His Ala Pro Arg Pro Arg Pro Ser Pro Phe Ser Glu
100          105          110
Trp Asp Gly Asn Ile Phe Lys Ile Pro Gln Val Leu His Ala Leu Xaa
115          120          125
His Tyr Asn Ala Arg His Pro Gly Gly Glu Phe Asp Val Val Lys
130          135          140
```

(2) INFORMATION FOR SEQ ID NO:189:

- (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 500 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- (ix) FEATURE:
 - (A) NAME/KEY: -
 - (B) LOCATION: 1..500
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481823

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:189:

agtccccact	ctcgcccccg	ctccctccaa	agtccaaacc	ctaccacccc	acttccccac	60
caccactaca	tggcggcggc	gctcgccctcc	tcccgctact	gctggagccg	cccgtcgctg	120
ccgccccaac	cgacccgcgg	ccgccgctcc	gtcactagct	gcgcgctctc	cggacgagag	180
aaaagaaact	ccttttagctg	gagagagtg	gcaatttctg	ttgcattgtc	agttggacta	240
atcactgggtg	caccaacggt	tggaccaccg	gcctatgctt	cttctcttga	acctgttctt	300
ccagatgtgt	ctgttcttat	ctctggacct	cccattaaag	atccaggtgc	tttattgaga	360
tatgctttac	caatagataa	taaagctatt	cggatgaagt	caaaagccgc	tggaggatat	420
cactgasagc	ctcaaggttg	stggkgttag	aggcttggat	tcagttgaaa	gaaaatgtca	480
gasaagcatc	gaaagcackg					

(2) INFORMATION FOR SEQ ID NO:190:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 131 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..131
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481824

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:190:

Ser	Pro	His	Ser	Arg	Pro	Arg	Ser	Leu	Gln	Ser	Pro	Asn	Pro	Thr	Thr
1			5					10						15	
Pro	Leu	Pro	His	His	His	Tyr	Met	Ala	Ala	Ala	Leu	Ala	Ser	Ser	Arg
			20					25					30		
Tyr	Cys	Trp	Ser	Arg	Pro	Ser	Leu	Pro	Pro	Gln	Pro	Thr	Arg	Gly	Arg
			35				40					45			
Arg	Ser	Val	Thr	Ser	Cys	Ala	Leu	Ser	Gly	Arg	Glu	Lys	Arg	Asn	Ser
			50			55					60				
Phe	Ser	Trp	Arg	Glu	Cys	Ala	Ile	Ser	Val	Ala	Leu	Ser	Val	Gly	Leu
65				70						75				80	
Ile	Thr	Gly	Ala	Pro	Thr	Phe	Gly	Pro	Pro	Ala	Tyr	Ala	Ser	Ser	Leu
			85					90					95		
Glu	Pro	Val	Leu	Pro	Asp	Val	Ser	Val	Leu	Ile	Ser	Gly	Pro	Pro	Ile
			100					105					110		
Lys	Asp	Pro	Gly	Ala	Leu	Leu	Arg	Tyr	Ala	Leu	Pro	Ile	Asp	Asn	Lys
			115				120						125		
Ala	Ile	Arg													
			130												

(2) INFORMATION FOR SEQ ID NO:191:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 79 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..79

(D) OTHER INFORMATION: / Ceres Seq. ID 1481825

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:191:

```
Val Pro Thr Leu Val Pro Ala Pro Ser Lys Val Gln Thr Leu Pro Pro
1          5          10          15
His Phe Pro Thr Thr Thr Trp Arg Arg Arg Ser Pro Pro Pro Ala
          20          25          30
Thr Ala Gly Ala Ala Arg Arg Cys Arg Pro Asn Arg Pro Ala Ala Ala
          35          40          45
Ala Pro Ser Leu Ala Ala Arg Ser Pro Asp Glu Arg Lys Glu Thr Pro
          50          55          60
Leu Ala Gly Glu Ser Val Gln Phe Leu Leu His Cys Gln Leu Asp
65          70          75
```

(2) INFORMATION FOR SEQ ID NO:192:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 108 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..108

(D) OTHER INFORMATION: / Ceres Seq. ID 1481826

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:192:

```
Met Ala Ala Ala Leu Ala Ser Ser Arg Tyr Cys Trp Ser Arg Pro Ser
1          5          10          15
Leu Pro Pro Gln Pro Thr Arg Gly Arg Arg Ser Val Thr Ser Cys Ala
          20          25          30
Leu Ser Gly Arg Glu Lys Arg Asn Ser Phe Ser Trp Arg Glu Cys Ala
          35          40          45
Ile Ser Val Ala Leu Ser Val Gly Leu Ile Thr Gly Ala Pro Thr Phe
          50          55          60
Gly Pro Pro Ala Tyr Ala Ser Ser Leu Glu Pro Val Leu Pro Asp Val
65          70          75          80
Ser Val Leu Ile Ser Gly Pro Pro Ile Lys Asp Pro Gly Ala Leu Leu
          85          90          95
Arg Tyr Ala Leu Pro Ile Asp Asn Lys Ala Ile Arg
          100          105
```

(2) INFORMATION FOR SEQ ID NO:193:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 876 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..876

(D) OTHER INFORMATION: / Ceres Seq. ID 1481827

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:193:

```
gaaggcacac tgccggcgct ctatcttccg ctctctctcc tactcgcgct cggcaaggcg      60
gcggcgcsst caggetcggg ctacggcgct acggcatccc ctccgcctct cgccagtcgc      120
ctcgcgccat cgctccgggt cacaccggcc ggtgactctg ctaaaaatgg tgttcctttc      180
aatgaagac gcctggatcc atgatgaaga catcatggat gatgttgatt cagatgttga      240
agaatcagac tcagaagggtg attcagggtga agaagctcag gctaagcctg cagacaaagc      300
gatatacaac aaggaggcta ttcttgaaaa acttgaagac atagcctggc ccaagaatgt      360
ggactggatg cacaaactca ctgttgagca tgatcaaggg gagaaagttg atgtgaacga      420
tgatcttgcc cgcgaacttg cgttttacac ccaagctttg gatggcacia ggcaggcctt      480
tgagaagctg cagtcgatga aggtccgggt cctcagacca acagattact acgctgagat      540
```

ggtgaagact gatgcacaca tgcacaagat caaggggagg ttgttgtcag agaagaagag 600
gattgaggaa gctgaggagc ggaggaaggc tagggagtcc aggaagaaaag caaaggaggt 660
tcaggctgag aagaagaagg agagggctaa gcagaagaag gagcagattg agtcagtcaa 720
gaagtggaga aagcagagac aacaaggggg attcaccaag ggaaatgatg atgtgccaga 780
ccttaatttt gaaggagaag aaggatttaa acaatcaaag aaaaagaggc ccggtgtttc 840
tcctggtgac aggtctggtg gtcttgcctt ctcttc

(2) INFORMATION FOR SEQ ID NO:194:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 291 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..291
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481828

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:194:

Lys Ala His Cys Arg Arg Ser Ile Phe Arg Ser Leu Ser Tyr Ser Arg
1 5 10 15
Ser Ala Arg Arg Arg Arg Xaa Gln Ala Arg Ala Thr Ala Ser Arg His
20 25 30
Pro Leu Arg Leu Ser Pro Val Ala Ser Arg His Arg Leu Arg Ser His
35 40 45
Arg Pro Val Thr Leu Leu Lys Met Val Phe Leu Ser Asn Glu Asp Ala
50 55 60
Trp Ile His Asp Glu Asp Ile Met Asp Asp Val Asp Ser Asp Val Glu
65 70 75 80
Glu Ser Asp Ser Glu Gly Asp Ser Gly Glu Glu Ala Gln Ala Lys Pro
85 90 95
Ala Asp Lys Ala Ile Tyr Asn Lys Glu Ala Ile Leu Glu Lys Leu Glu
100 105 110
Asp Ile Ala Trp Pro Lys Asn Val Asp Trp Met His Lys Leu Thr Val
115 120 125
Glu His Asp Gln Gly Glu Lys Val Asp Val Asn Asp Asp Leu Ala Arg
130 135 140
Glu Leu Ala Phe Tyr Thr Gln Ala Leu Asp Gly Thr Arg Gln Ala Phe
145 150 155 160
Glu Lys Leu Gln Ser Met Lys Val Arg Phe Leu Arg Pro Thr Asp Tyr
165 170 175
Tyr Ala Glu Met Val Lys Thr Asp Ala His Met His Lys Ile Lys Gly
180 185 190
Arg Leu Leu Ser Glu Lys Lys Arg Ile Glu Glu Ala Glu Glu Arg Arg
195 200 205
Lys Ala Arg Glu Ser Arg Lys Lys Ala Lys Glu Val Gln Ala Glu Lys
210 215 220
Lys Lys Glu Arg Ala Lys Gln Lys Lys Glu Gln Ile Glu Ser Val Lys
225 230 235 240
Lys Trp Arg Lys Gln Arg Gln Gln Gly Gly Phe Thr Lys Gly Asn Asp
245 250 255
Asp Val Pro Asp Leu Asn Phe Glu Gly Glu Glu Gly Phe Lys Gln Ser
260 265 270
Lys Lys Lys Arg Pro Gly Val Ser Pro Gly Asp Arg Ser Gly Gly Leu
275 280 285
Ala Phe Ser
290

(2) INFORMATION FOR SEQ ID NO:195:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 236 amino acids

- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..236
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481829
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:195:

Met	Val	Phe	Leu	Ser	Asn	Glu	Asp	Ala	Trp	Ile	His	Asp	Glu	Asp	Ile
1				5				10					15		
Met	Asp	Asp	Val	Asp	Ser	Asp	Val	Glu	Glu	Ser	Asp	Ser	Glu	Gly	Asp
			20					25					30		
Ser	Gly	Glu	Glu	Ala	Gln	Ala	Lys	Pro	Ala	Asp	Lys	Ala	Ile	Tyr	Asn
		35					40					45			
Lys	Glu	Ala	Ile	Leu	Glu	Lys	Leu	Glu	Asp	Ile	Ala	Trp	Pro	Lys	Asn
	50					55				60					
Val	Asp	Trp	Met	His	Lys	Leu	Thr	Val	Glu	His	Asp	Gln	Gly	Glu	Lys
65					70				75					80	
Val	Asp	Val	Asn	Asp	Asp	Leu	Ala	Arg	Glu	Leu	Ala	Phe	Tyr	Thr	Gln
			85					90					95		
Ala	Leu	Asp	Gly	Thr	Arg	Gln	Ala	Phe	Glu	Lys	Leu	Gln	Ser	Met	Lys
			100				105					110			
Val	Arg	Phe	Leu	Arg	Pro	Thr	Asp	Tyr	Tyr	Ala	Glu	Met	Val	Lys	Thr
		115				120					125				
Asp	Ala	His	Met	His	Lys	Ile	Lys	Gly	Arg	Leu	Leu	Ser	Glu	Lys	Lys
	130				135				140						
Arg	Ile	Glu	Glu	Ala	Glu	Glu	Arg	Arg	Lys	Ala	Arg	Glu	Ser	Arg	Lys
145					150				155					160	
Lys	Ala	Lys	Glu	Val	Gln	Ala	Glu	Lys	Lys	Lys	Glu	Arg	Ala	Lys	Gln
			165					170					175		
Lys	Lys	Glu	Gln	Ile	Glu	Ser	Val	Lys	Lys	Trp	Arg	Lys	Gln	Arg	Gln
		180					185					190			
Gln	Gly	Gly	Phe	Thr	Lys	Gly	Asn	Asp	Asp	Val	Pro	Asp	Leu	Asn	Phe
		195				200					205				
Glu	Gly	Glu	Glu	Gly	Phe	Lys	Gln	Ser	Lys	Lys	Lys	Arg	Pro	Gly	Val
	210					215					220				
Ser	Pro	Gly	Asp	Arg	Ser	Gly	Gly	Leu	Ala	Phe	Ser				
225					230				235						

- (2) INFORMATION FOR SEQ ID NO:196:
 - (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 220 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
 - (ii) MOLECULE TYPE: peptide
 - (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..220
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481830
 - (xi) SEQUENCE DESCRIPTION: SEQ ID NO:196:

Met	Asp	Asp	Val	Asp	Ser	Asp	Val	Glu	Glu	Ser	Asp	Ser	Glu	Gly	Asp
1			5					10					15		
Ser	Gly	Glu	Glu	Ala	Gln	Ala	Lys	Pro	Ala	Asp	Lys	Ala	Ile	Tyr	Asn
			20				25					30			
Lys	Glu	Ala	Ile	Leu	Glu	Lys	Leu	Glu	Asp	Ile	Ala	Trp	Pro	Lys	Asn
	35					40					45				
Val	Asp	Trp	Met	His	Lys	Leu	Thr	Val	Glu	His	Asp	Gln	Gly	Glu	Lys
	50				55						60				

Val Asp Val Asn Asp Asp Leu Ala Arg Glu Leu Ala Phe Tyr Thr Gln
65 70 75 80
Ala Leu Asp Gly Thr Arg Gln Ala Phe Glu Lys Leu Gln Ser Met Lys
85 90 95
Val Arg Phe Leu Arg Pro Thr Asp Tyr Tyr Ala Glu Met Val Lys Thr
100 105 110
Asp Ala His Met His Lys Ile Lys Gly Arg Leu Leu Ser Glu Lys Lys
115 120 125
Arg Ile Glu Glu Ala Glu Glu Arg Arg Lys Ala Arg Glu Ser Arg Lys
130 135 140
Lys Ala Lys Glu Val Gln Ala Glu Lys Lys Lys Glu Arg Ala Lys Gln
145 150 155 160
Lys Lys Glu Gln Ile Glu Ser Val Lys Lys Trp Arg Lys Gln Arg Gln
165 170 175
Gln Gly Gly Phe Thr Lys Gly Asn Asp Val Pro Asp Leu Asn Phe
180 185 190
Glu Gly Glu Glu Gly Phe Lys Gln Ser Lys Lys Lys Arg Pro Gly Val
195 200 205
Ser Pro Gly Asp Arg Ser Gly Gly Leu Ala Phe Ser
210 215 220

(2) INFORMATION FOR SEQ ID NO:197:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 530 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..530
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481831

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:197:

aagctcgtct	cgcaccagaa	acccgcgaat	caatcccga	tcccgatcga	ccggcggcgc	60
gggaggcgat	gccggacaag	gcggtggacg	acgtcatgga	ggccgcggtg	ggggcccact	120
tcagcggcct	cgccctcgag	gcgctgcgcc	tctccacctc	tgcgccctct	tccccttctt	180
cctccccgcg	cgcgcggcgc	cacacgcact	ccaacggagc	cgtctacgcc	aacggcacca	240
ccgagcttcc	ctctcctgcc	gctgcccggc	agccattcgt	catcgggggtt	tctggaggga	300
cggcgtcggg	gaagacgacg	gtgtgcgaca	tgatcatcca	gcagctgcac	gaccaccgtg	360
tcggtgctcg	taaccaggat	tcgttttacc	gtggccttac	tgctgaagag	tctgcacacg	420
cacaagacta	taattttgat	caccctgatg	catttgatac	agagcaactt	ctagaatgca	480
tggggcagct	gaaatgtgct	caacctgtaa	atgttcctat	atatgatttc		

(2) INFORMATION FOR SEQ ID NO:198:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 176 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..176
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481832

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:198:

Ala Arg Leu Ala Pro Glu Thr Arg Glu Ser Ile Pro Asn Pro Asp Arg
1 5 10 15
Pro Ala Ala Arg Glu Ala Met Pro Asp Lys Ala Val Asp Asp Val Met
20 25 30
Glu Ala Ala Val Gly Ala His Phe Ser Gly Leu Arg Leu Glu Ala Leu
35 40 45

Arg	Leu	Ser	Thr	Ser	Ala	Pro	Ser	Ser	Pro	Ser	Ser	Ser	Pro	Ala	Ala
50						55					60				
Ala	Ala	His	Thr	His	Ser	Asn	Gly	Ala	Val	Tyr	Ala	Asn	Gly	Thr	Thr
65					70					75				80	
Glu	Leu	Pro	Ser	Pro	Ala	Ala	Ala	Arg	Gln	Pro	Phe	Val	Ile	Gly	Val
				85					90					95	
Ser	Gly	Gly	Thr	Ala	Ser	Gly	Lys	Thr	Thr	Val	Cys	Asp	Met	Ile	Ile
			100					105					110		
Gln	Gln	Leu	His	Asp	His	Arg	Val	Val	Leu	Val	Asn	Gln	Asp	Ser	Phe
		115					120					125			
Tyr	Arg	Gly	Leu	Thr	Ala	Glu	Glu	Ser	Ala	His	Ala	Gln	Asp	Tyr	Asn
130						135					140				
Phe	Asp	His	Pro	Asp	Ala	Phe	Asp	Thr	Glu	Gln	Leu	Leu	Glu	Cys	Met
145					150					155				160	
Gly	Gln	Leu	Lys	Cys	Ala	Gln	Pro	Val	Asn	Val	Pro	Ile	Tyr	Asp	Phe
				165					170					175	

(2) INFORMATION FOR SEQ ID NO:199:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 154 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..154
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481833

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:199:

Met	Pro	Asp	Lys	Ala	Val	Asp	Asp	Val	Met	Glu	Ala	Ala	Val	Gly	Ala
1			5					10						15	
His	Phe	Ser	Gly	Leu	Arg	Leu	Glu	Ala	Leu	Arg	Leu	Ser	Thr	Ser	Ala
			20				25						30		
Pro	Ser	Ser	Pro	Ser	Ser	Ser	Pro	Ala	Ala	Ala	Ala	His	Thr	His	Ser
			35				40					45			
Asn	Gly	Ala	Val	Tyr	Ala	Asn	Gly	Thr	Thr	Glu	Leu	Pro	Ser	Pro	Ala
50						55				60					
Ala	Ala	Arg	Gln	Pro	Phe	Val	Ile	Gly	Val	Ser	Gly	Gly	Thr	Ala	Ser
65				70					75					80	
Gly	Lys	Thr	Thr	Val	Cys	Asp	Met	Ile	Ile	Gln	Gln	Leu	His	Asp	His
			85					90					95		
Arg	Val	Val	Leu	Val	Asn	Gln	Asp	Ser	Phe	Tyr	Arg	Gly	Leu	Thr	Ala
			100					105					110		
Glu	Glu	Ser	Ala	His	Ala	Gln	Asp	Tyr	Asn	Phe	Asp	His	Pro	Asp	Ala
		115					120					125			
Phe	Asp	Thr	Glu	Gln	Leu	Leu	Glu	Cys	Met	Gly	Gln	Leu	Lys	Cys	Ala
130					135					140					
Gln	Pro	Val	Asn	Val	Pro	Ile	Tyr	Asp	Phe						
145				150											

(2) INFORMATION FOR SEQ ID NO:200:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 145 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide

(B) LOCATION: 1..145

(D) OTHER INFORMATION: / Ceres Seq. ID 1481834

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:200:

```
Met Glu Ala Ala Val Gly Ala His Phe Ser Gly Leu Arg Leu Glu Ala
1      5      10      15
Leu Arg Leu Ser Thr Ser Ala Pro Ser Ser Pro Ser Ser Pro Ala
      20      25      30
Ala Ala Ala His Thr His Ser Asn Gly Ala Val Tyr Ala Asn Gly Thr
      35      40      45
Thr Glu Leu Pro Ser Pro Ala Ala Ala Arg Gln Pro Phe Val Ile Gly
      50      55      60
Val Ser Gly Gly Thr Ala Ser Gly Lys Thr Thr Val Cys Asp Met Ile
65      70      75      80
Ile Gln Gln Leu His Asp His Arg Val Val Leu Val Asn Gln Asp Ser
      85      90      95
Phe Tyr Arg Gly Leu Thr Ala Glu Glu Ser Ala His Ala Gln Asp Tyr
      100     105     110
Asn Phe Asp His Pro Asp Ala Phe Asp Thr Glu Gln Leu Leu Glu Cys
      115     120     125
Met Gly Gln Leu Lys Cys Ala Gln Pro Val Asn Val Pro Ile Tyr Asp
130     135     140
```

Phe
145

(2) INFORMATION FOR SEQ ID NO:201:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1087 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1087
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481839

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:201:

```
aatcaagctg actcggttgc tccgccgtgt taccgtccag ctcaaggaac gattccaagc      60
tcgatatactg ctgttttcaga tcataatccg gctcaaatcc gtaactcagg ttctgaaact      120
cgtccccatt tccaaatcct atacactcca ccgggaaact cggctgactt aactcgctct      180
cactctgaga gacaacgcag accttcttcg ccggaggctg atcggttgat cgtggcgatg      240
acggcggagg agtataatta ggaggaggag gatggtgcag atctgggaag ttgagcttgg      300
ctttatcacc acggatctgc ttggccgcaa catcataagc catggcagct tcctccgccg      360
tgttgaacgt accaagccaa actctaacac cttttcgtgg atctcgaatc tcagccgcc      420
attttcccca tggacgctta cgtatccctc tataaacatt cttcctcttc ctccgtttcc      480
ccggtctctgt tgctgtctcc ttcttctactg cctcctcttt cacgttaact tcaaaatttt      540
cacccgattc cccaaagtcc aaaattttaca attttaacct cacacagata attaaataat      600
cctgataaat tacattacca aaaccacaaa tatttttttt ttatcatctt ccgtaagtgc      660
cagaaatatt attttacctt tttgctaaaa aggttagaaa aaactatatg tttgtgtttt      720
tgaatgattt tgtatttttg tttatgattm ataggagagt acataccttg gttggtggga      780
tggagtttgg aggtggaata gaaaccccag aagtcgtcgg cggcggaagc atcgagctct      840
gaccagagtt cctcagccgt gagtttacgg cccttggcct tggtgacgag aggggcataa      900
tcggaaataa tagcaccgcc acacattttc tctgtttggt gctgtgggtt tctttcaaga      960
gaaagtttcc tacggtggag ctgaaatgcc tttataggcg caaaataaat gttttatggt      1020
aataaagtgt gagtgaaatg aattacttta tattagaata ataattctaa tagttttatg      1080
ttccttg
```

(2) INFORMATION FOR SEQ ID NO:202:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 86 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:

(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..86
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481840
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:202:
Asn Gln Ala Asp Ser Val Ala Pro Pro Cys Tyr Arg Pro Ala Gln Gly
1 5 10 15
Thr Ile Pro Ser Ser Ile Ser Ala Val Ser Asp His Asn Pro Ala Gln
 20 25 30
Ile Arg Asn Ser Gly Ser Glu Thr Arg Pro His Phe Gln Ile Leu Tyr
 35 40 45
Thr Pro Pro Gly Asn Ser Ala Asp Leu Thr Arg Ser His Ser Glu Arg
 50 55 60
Gln Arg Arg Pro Ser Ser Pro Glu Ala Asp Arg Leu Ile Val Ala Met
65 70 75 80
Thr Ala Glu Glu Tyr Asn
 85
(2) INFORMATION FOR SEQ ID NO:203:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 56 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: peptide
 (ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..56
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481841
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:203:
Ile Lys Leu Thr Arg Leu Leu Arg Arg Val Thr Val Gln Leu Lys Glu
1 5 10 15
Arg Phe Gln Ala Arg Tyr Leu Leu Phe Gln Ile Ile Ile Arg Leu Lys
 20 25 30
Ser Val Thr Gln Val Leu Lys Leu Val Pro Ile Ser Lys Ser Tyr Thr
 35 40 45
Leu His Arg Glu Thr Arg Leu Thr
 50 55
(2) INFORMATION FOR SEQ ID NO:204:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 101 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: peptide
 (ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..101
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481842
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:204:
Met Val Gln Ile Trp Glu Val Glu Leu Gly Phe Ile Thr Thr Asp Leu
1 5 10 15
Leu Gly Arg Asn Ile Ile Ser His Gly Ser Phe Leu Arg Arg Val Glu
 20 25 30
Arg Thr Lys Pro Asn Ser Asn Thr Phe Ser Trp Ile Ser Asn Leu Ser
 35 40 45
Arg Pro Phe Ser Pro Trp Thr Leu Thr Tyr Pro Ser Ile Asn Ile Leu
 50 55 60

Pro Leu Pro Pro Phe Pro Arg Leu Cys Cys Leu Leu Leu Leu His Cys
65 70 75 80
Leu Leu Phe His Val Asn Phe Lys Ile Phe Thr Gly Phe Pro Lys Val
85 90 95
Gln Asn Leu Gln Phe
100

(2) INFORMATION FOR SEQ ID NO:205:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1160 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1160
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481847

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:205:

gaggtttctt	gggaacagga	tcgcttctac	agatatattca	cccatgtcaa	gtttaacagt	60
gagaagggtta	tcgcgcgaag	atatacaact	ggttcagaat	ctaattgaac	gatgcctcca	120
gctttacatg	aaccagaaaag	aagttgttga	cactcttcta	gaacaggcta	agatcgaacc	180
tggtttttaca	gaactagttt	ggcagaagct	tgaagaagag	aaccgcgaat	ttttcaaggc	240
atattatctg	aggctcatgg	tgaagcacca	gataatggaa	tataacgaac	tgcttgagca	300
gcagataaac	cacatgcgcc	agatgcatcc	aactgcaggg	gcttctgttc	gaaacaggaa	360
tggttctcat	gttccaccaa	tgaatcagca	acaattactc	tatgaacgca	aggaaccaga	420
tcaatcctct	cctaattctgt	caagtccata	cctcaatgga	ggctcagcaa	ttaacacaaa	480
tataccttct	tatgtggact	tttcatccca	ttctagaaga	ggtgatcctt	caccaaactc	540
gctctccttg	caggccacaa	atatgccttt	gatgcaagga	atgatcaagt	ctgagactgc	600
atatcaaaac	tgtgctccat	acatgtatgg	tggtgaagca	cagtccacag	ttggagatgt	660
caccatcgca	tctttcagca	atgattccag	caaccaatcc	ctgaatgato	ctcttgctga	720
tccagatgct	cctacatttg	gctcgtagg	acaaattcct	cagaacttca	gcctctctga	780
tctgacagct	gatttttccc	agagttcaga	tattctggag	agctacgagg	gatcaccggt	840
cctattggct	gatgctgaaa	atttcctgga	ctctagcgaa	agggtagaac	atcaaggaga	900
ccacgaaaga	ttgaggacca	tatcatcagg	cttcagttac	gaaaacttcc	gaagcaatta	960
ggttttattac	acatggaact	tcgtagtcat	gctttttacgt	ctgcaactac	ttgcaggatt	1020
taatcccatg	atcagtgtac	atagatattc	ttacctttcg	aaagacaatt	ttgggggttca	1080
gggtgattac	taatattatt	attctcaagt	gtagagaaat	ttgggttttta	gtaataaata	1140
tttaagaacc	tgttgatttt					

(2) INFORMATION FOR SEQ ID NO:206:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 319 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..319
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481848

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:206:

Arg	Phe	Leu	Gly	Asn	Arg	Ile	Ala	Ser	Thr	Asp	Ile	Ser	Pro	Met	Ser
1				5					10					15	
Ser	Leu	Thr	Val	Arg	Arg	Val	Ser	Arg	Glu	Asp	Ile	Gln	Leu	Val	Gln
			20					25					30		
Asn	Leu	Ile	Glu	Arg	Cys	Leu	Gln	Leu	Tyr	Met	Asn	Gln	Lys	Glu	Val
		35					40					45			
Val	Asp	Thr	Leu	Leu	Glu	Gln	Ala	Lys	Ile	Glu	Pro	Gly	Phe	Thr	Glu
	50					55					60				
Leu	Val	Trp	Gln	Lys	Leu	Glu	Glu	Asn	Arg	Glu	Phe	Phe	Lys	Ala	

65				70					75				80
Tyr	Tyr	Leu	Arg	Leu	Met	Val	Lys	His	Gln	Ile	Met	Glu	Tyr
				85					90				95
Leu	Leu	Glu	Gln	Gln	Ile	Asn	His	Met	Arg	Gln	Met	His	Pro
				100					105				110
Gly	Ala	Ser	Val	Arg	Asn	Arg	Asn	Gly	Ser	His	Val	Pro	Met
				115					120				125
Gln	Gln	Gln	Leu	Leu	Tyr	Glu	Arg	Lys	Glu	Pro	Asp	Gln	Ser
				130					135				140
Asn	Leu	Ser	Ser	Pro	Tyr	Leu	Asn	Gly	Gly	Ser	Ala	Ile	Asn
				145					150				155
Ile	Pro	Ser	Tyr	Val	Asp	Phe	Ser	Ser	His	Ser	Arg	Arg	Val
				165					170				175
Ser	Pro	Asn	Ser	Leu	Ser	Leu	Gln	Ala	Thr	Asn	Met	Pro	Leu
				180					185				190
Gly	Met	Ile	Lys	Ser	Glu	Thr	Ala	Tyr	Gln	Asn	Cys	Ala	Pro
				195					200				205
Tyr	Gly	Gly	Glu	Ala	Gln	Ser	Thr	Val	Gly	Asp	Val	Thr	Ile
				210					215				220
Phe	Ser	Asn	Asp	Ser	Ser	Asn	Gln	Ser	Leu	Asn	Asp	Pro	Leu
				225					230				235
Pro	Asp	Ala	Pro	Thr	Phe	Gly	Ser	Leu	Gly	Gln	Ile	Pro	Gln
				245					250				255
Ser	Leu	Ser	Asp	Leu	Thr	Ala	Asp	Phe	Ser	Gln	Ser	Ser	Asp
				260					265				270
Glu	Ser	Tyr	Glu	Gly	Ser	Pro	Phe	Leu	Leu	Ala	Asp	Ala	Glu
				275					280				285
Leu	Asp	Ser	Ser	Glu	Arg	Val	Glu	His	Gln	Gly	Asp	His	Glu
				290					295				300
Arg	Thr	Ile	Ser	Ser	Gly	Phe	Ser	Tyr	Glu	Asn	Phe	Arg	Ser
				305					310				315

(2) INFORMATION FOR SEQ ID NO:207:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 305 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..305

(D) OTHER INFORMATION: / Ceres Seq. ID 1481849

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:207:

Met	Ser	Ser	Leu	Thr	Val	Arg	Arg	Val	Ser	Arg	Glu	Asp	Ile	Gln	Leu
1			5					10						15	
Val	Gln	Asn	Leu	Ile	Glu	Arg	Cys	Leu	Gln	Leu	Tyr	Met	Asn	Gln	Lys
			20					25						30	
Glu	Val	Val	Asp	Thr	Leu	Leu	Glu	Gln	Ala	Lys	Ile	Glu	Pro	Gly	Phe
			35					40						45	
Thr	Glu	Leu	Val	Trp	Gln	Lys	Leu	Glu	Glu	Glu	Asn	Arg	Glu	Phe	Phe
			50				55				60				
Lys	Ala	Tyr	Tyr	Leu	Arg	Leu	Met	Val	Lys	His	Gln	Ile	Met	Glu	Tyr
				65			70				75			80	
Asn	Glu	Leu	Leu	Glu	Gln	Gln	Ile	Asn	His	Met	Arg	Gln	Met	His	Pro
				85				90						95	
Thr	Ala	Gly	Ala	Ser	Val	Arg	Asn	Arg	Asn	Gly	Ser	His	Val	Pro	Pro
			100					105						110	
Met	Asn	Gln	Gln	Leu	Leu	Tyr	Glu	Arg	Lys	Glu	Pro	Asp	Gln	Ser	
			115					120						125	

Ser	Pro	Asn	Leu	Ser	Ser	Pro	Tyr	Leu	Asn	Gly	Gly	Ser	Ala	Ile	Asn
130						135					140				
Thr	Asn	Ile	Pro	Ser	Tyr	Val	Asp	Phe	Ser	Ser	His	Ser	Arg	Arg	Val
145					150					155					160
Asp	Pro	Ser	Pro	Asn	Ser	Leu	Ser	Leu	Gln	Ala	Thr	Asn	Met	Pro	Leu
				165					170					175	
Met	Gln	Gly	Met	Ile	Lys	Ser	Glu	Thr	Ala	Tyr	Gln	Asn	Cys	Ala	Pro
			180					185					190		
Tyr	Met	Tyr	Gly	Gly	Glu	Ala	Gln	Ser	Thr	Val	Gly	Asp	Val	Thr	Ile
		195					200					205			
Ala	Ser	Phe	Ser	Asn	Asp	Ser	Ser	Asn	Gln	Ser	Leu	Asn	Asp	Pro	Leu
	210					215					220				
Val	Asp	Pro	Asp	Ala	Pro	Thr	Phe	Gly	Ser	Leu	Gly	Gln	Ile	Pro	Gln
225					230					235					240
Asn	Phe	Ser	Leu	Ser	Asp	Leu	Thr	Ala	Asp	Phe	Ser	Gln	Ser	Ser	Asp
				245					250					255	
Ile	Leu	Glu	Ser	Tyr	Glu	Gly	Ser	Pro	Phe	Leu	Leu	Ala	Asp	Ala	Glu
			260					265					270		
Asn	Phe	Leu	Asp	Ser	Ser	Glu	Arg	Val	Glu	His	Gln	Gly	Asp	His	Glu
		275					280					285			
Arg	Leu	Arg	Thr	Ile	Ser	Ser	Gly	Phe	Ser	Tyr	Glu	Asn	Phe	Arg	Ser
	290					295					300				

Asn
305

(2) INFORMATION FOR SEQ ID NO:208:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 277 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..277

(D) OTHER INFORMATION: / Ceres Seq. ID 1481850

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:208:

Met	Asn	Gln	Lys	Glu	Val	Val	Asp	Thr	Leu	Leu	Glu	Gln	Ala	Lys	Ile
1			5						10					15	
Glu	Pro	Gly	Phe	Thr	Glu	Leu	Val	Trp	Gln	Lys	Leu	Glu	Glu	Glu	Asn
			20					25					30		
Arg	Glu	Phe	Lys	Ala	Tyr	Tyr	Leu	Arg	Leu	Met	Val	Lys	His	Gln	
	35				40						45				
Ile	Met	Glu	Tyr	Asn	Glu	Leu	Leu	Glu	Gln	Gln	Ile	Asn	His	Met	Arg
	50					55					60				
Gln	Met	His	Pro	Thr	Ala	Gly	Ala	Ser	Val	Arg	Asn	Arg	Asn	Gly	Ser
65					70					75					80
His	Val	Pro	Pro	Met	Asn	Gln	Gln	Gln	Leu	Leu	Tyr	Glu	Arg	Lys	Glu
				85					90					95	
Pro	Asp	Gln	Ser	Ser	Pro	Asn	Leu	Ser	Ser	Pro	Tyr	Leu	Asn	Gly	Gly
			100					105					110		
Ser	Ala	Ile	Asn	Thr	Asn	Ile	Pro	Ser	Tyr	Val	Asp	Phe	Ser	Ser	His
		115					120					125			
Ser	Arg	Arg	Val	Asp	Pro	Ser	Pro	Asn	Ser	Leu	Ser	Leu	Gln	Ala	Thr
	130					135					140				
Asn	Met	Pro	Leu	Met	Gln	Gly	Met	Ile	Lys	Ser	Glu	Thr	Ala	Tyr	Gln
145					150					155					160
Asn	Cys	Ala	Pro	Tyr	Met	Tyr	Gly	Gly	Glu	Ala	Gln	Ser	Thr	Val	Gly
				165					170					175	
Asp	Val	Thr	Ile	Ala	Ser	Phe	Ser	Asn	Asp	Ser	Ser	Asn	Gln	Ser	Leu

180 185 190
Asn Asp Pro Leu Val Asp Pro Asp Ala Pro Thr Phe Gly Ser Leu Gly
195 200 205
Gln Ile Pro Gln Asn Phe Ser Leu Ser Asp Leu Thr Ala Asp Phe Ser
210 215 220
Gln Ser Ser Asp Ile Leu Glu Ser Tyr Glu Gly Ser Pro Phe Leu Leu
225 230 235 240
Ala Asp Ala Glu Asn Phe Leu Asp Ser Ser Glu Arg Val Glu His Gln
245 250 255
Gly Asp His Glu Arg Leu Arg Thr Ile Ser Ser Gly Phe Ser Tyr Glu
260 265 270
Asn Phe Arg Ser Asn
275

(2) INFORMATION FOR SEQ ID NO:209:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 806 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..806
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481851

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:209:

cagcttaaca	cttgatgagg	ttcaaaatca	cttggggagt	tctggtaaag	ctctgggaag	60
catgaacctt	gatgagcttt	tgaagagtgt	ctgttctgtt	gaagctaata	agccatcgtc	120
tatggctgtc	aatgggtggag	cagctgctca	ggagggtctt	tctcgccagg	ggagtttgac	180
tttgccctcg	gatctcagca	aaaagactgt	tgatgagggt	tggaaagaca	ttcagcagaa	240
taagaatgga	ggtagtgtct	atgagaggag	ggataagcag	cctacacttg	gggaaatgac	300
gcttgaagac	ctgttggtga	aagcaggagt	ggtcactgag	actatccctg	gttcgaacca	360
tgatggctct	gttggtggtg	gtagtgtctg	ttcagggtgt	ggtttagggc	aaaacattac	420
tcaagttggc	ccatggattc	aatatcatca	gctcccatca	atgccacagc	ctcaagcatt	480
tatgccctat	ccggtttcag	atatgcaagc	aatgggtgtc	cagtcttctt	tgatgggtgg	540
tttgtcagat	acacaaactc	ctggaaggaa	gagggtagct	tcaggagaag	ttgtagagaa	600
gactgtgaca	ccattgcttg	catagctgca	acaggtaaag	gtccactcaa	caattgggct	660
actcacctca	gtgatccact	ccacaccacc	atcatcgata	ccttctctct	atcttaaaat	720
cattatcatg	tgagattcta	tttgtaactt	atgtaaaaac	agagctatga	tgatactgaa	780
tcgactttgg	gcttttgctt	gtttgg				

(2) INFORMATION FOR SEQ ID NO:210:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 207 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..207
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481852

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:210:

Ser	Leu	Thr	Leu	Asp	Glu	Val	Gln	Asn	His	Leu	Gly	Ser	Ser	Gly	Lys
1				5				10				15			
Ala	Leu	Gly	Ser	Met	Asn	Leu	Asp	Glu	Leu	Leu	Lys	Ser	Val	Cys	Ser
		20						25				30			
Val	Glu	Ala	Asn	Gln	Pro	Ser	Ser	Met	Ala	Val	Asn	Gly	Gly	Ala	Ala
		35						40				45			
Ala	Gln	Glu	Gly	Leu	Ser	Arg	Gln	Gly	Ser	Leu	Thr	Leu	Pro	Arg	Asp
		50						55				60			

Leu Ser Lys Lys Thr Val Asp Glu Val Trp Lys Asp Ile Gln Gln Asn
65 70 75 80
Lys Asn Gly Gly Ser Ala His Glu Arg Arg Asp Lys Gln Pro Thr Leu
85 90 95
Gly Glu Met Thr Leu Glu Asp Leu Leu Lys Ala Gly Val Val Thr
100 105 110
Glu Thr Ile Pro Gly Ser Asn His Asp Gly Pro Val Gly Gly Ser
115 120 125
Ala Gly Ser Gly Ala Gly Leu Gly Gln Asn Ile Thr Gln Val Gly Pro
130 135 140
Trp Ile Gln Tyr His Gln Leu Pro Ser Met Pro Gln Pro Gln Ala Phe
145 150 155 160
Met Pro Tyr Pro Val Ser Asp Met Gln Ala Met Val Ser Gln Ser Ser
165 170 175
Leu Met Gly Gly Leu Ser Asp Thr Gln Thr Pro Gly Arg Lys Arg Val
180 185 190
Ala Ser Gly Glu Val Val Glu Lys Thr Val Thr Pro Leu Leu Ala
195 200 205

(2) INFORMATION FOR SEQ ID NO:211:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 187 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..187
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481853

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:211:

Met Asn Leu Asp Glu Leu Leu Lys Ser Val Cys Ser Val Glu Ala Asn
1 5 10 15
Gln Pro Ser Ser Met Ala Val Asn Gly Gly Ala Ala Ala Gln Glu Gly
20 25 30
Leu Ser Arg Gln Gly Ser Leu Thr Leu Pro Arg Asp Leu Ser Lys Lys
35 40 45
Thr Val Asp Glu Val Trp Lys Asp Ile Gln Gln Asn Lys Asn Gly Gly
50 55 60
Ser Ala His Glu Arg Arg Asp Lys Gln Pro Thr Leu Gly Glu Met Thr
65 70 75 80
Leu Glu Asp Leu Leu Lys Ala Gly Val Val Thr Glu Thr Ile Pro
85 90 95
Gly Ser Asn His Asp Gly Pro Val Gly Gly Gly Ser Ala Gly Ser Gly
100 105 110
Ala Gly Leu Gly Gln Asn Ile Thr Gln Val Gly Pro Trp Ile Gln Tyr
115 120 125
His Gln Leu Pro Ser Met Pro Gln Pro Gln Ala Phe Met Pro Tyr Pro
130 135 140
Val Ser Asp Met Gln Ala Met Val Ser Gln Ser Ser Leu Met Gly Gly
145 150 155 160
Leu Ser Asp Thr Gln Thr Pro Gly Arg Lys Arg Val Ala Ser Gly Glu
165 170 175
Val Val Glu Lys Thr Val Thr Pro Leu Leu Ala
180 185

(2) INFORMATION FOR SEQ ID NO:212:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 167 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:

- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..167
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481854

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:212:

Met	Ala	Val	Asn	Gly	Gly	Ala	Ala	Ala	Gln	Glu	Gly	Leu	Ser	Arg	Gln	
1				5					10					15		
Gly	Ser	Leu	Thr	Leu	Pro	Arg	Asp	Leu	Ser	Lys	Lys	Thr	Val	Asp	Glu	
			20					25					30			
Val	Trp	Lys	Asp	Ile	Gln	Gln	Asn	Lys	Asn	Gly	Gly	Ser	Ala	His	Glu	
			35				40					45				
Arg	Arg	Asp	Lys	Gln	Pro	Thr	Leu	Gly	Glu	Met	Thr	Leu	Glu	Asp	Leu	
			50				55				60					
Leu	Leu	Lys	Ala	Gly	Val	Val	Thr	Glu	Thr	Ile	Pro	Gly	Ser	Asn	His	
65					70				75					80		
Asp	Gly	Pro	Val	Gly	Gly	Gly	Ser	Ala	Gly	Ser	Gly	Ala	Gly	Leu	Gly	
				85					90					95		
Gln	Asn	Ile	Thr	Gln	Val	Gly	Pro	Trp	Ile	Gln	Tyr	His	Gln	Leu	Pro	
			100					105					110			
Ser	Met	Pro	Gln	Pro	Gln	Ala	Phe	Met	Pro	Tyr	Pro	Val	Ser	Asp	Met	
			115				120					125				
Gln	Ala	Met	Val	Ser	Gln	Ser	Ser	Leu	Met	Gly	Gly	Leu	Ser	Asp	Thr	
			130				135					140				
Gln	Thr	Pro	Gly	Arg	Lys	Arg	Val	Ala	Ser	Gly	Glu	Val	Val	Glu	Lys	
145					150					155					160	
Thr	Val	Thr	Pro	Leu	Leu	Ala										
					165											

(2) INFORMATION FOR SEQ ID NO:213:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 391 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- (ix) FEATURE:
 - (A) NAME/KEY: -
 - (B) LOCATION: 1..391
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481859

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:213:

acatatgctg	tccgtcaccg	cgcgcgcctc	cttgcccttc	ttcacccttt	cctcccggac	60
ccggcgtctc	cgtgcccgtg	cgttccttct	ccggcctgcg	gtctcctcca	ccggccaccg	120
ctgcctcgca	attgggcaag	gcaatcagac	cgtcctccatc	aaccgccctg	accgcgcgag	180
gaagatcaaa	cgcggaggcc	gtccgtccct	cccgcgttat	cgtgatgcca	ggcctcgccg	240
cagccgagca	ggacgccgtc	tgcgtggtgc	ggcgcgtcgc	ccgsgctctc	aaccgccgct	300
tcaccgacat	cgtcgcactg	ctcttcagcc	acaagggcgc	tggatcgctc	ggcgcmgtcg	360
cggggttcgc	matcgccgtc	gtgttcgcgt	g			

(2) INFORMATION FOR SEQ ID NO:214:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 130 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..130
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481860

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:214:

His	Met	Leu	Ser	Val	Thr	Ala	Arg	Ala	Ser	Leu	Pro	Phe	Phe	Thr	Leu
1				5					10					15	
Ser	Ser	Arg	Thr	Arg	Arg	Leu	Arg	Ala	Arg	Ala	Xaa	Leu	Leu	Arg	Pro
			20					25					30		
Ala	Val	Ser	Ser	Thr	Gly	His	Arg	Cys	Leu	Ala	Ile	Gly	Gln	Gly	Asn
		35				40						45			
Gln	Thr	Ala	Pro	Ile	Asn	Arg	Pro	Asp	Arg	Ala	Arg	Lys	Ile	Lys	Arg
	50				55					60					
Gly	Gly	Arg	Pro	Ser	Leu	Pro	Arg	Tyr	Arg	Asp	Ala	Arg	Pro	Arg	Arg
65				70				75						80	
Ser	Arg	Ala	Gly	Arg	Arg	Leu	Ala	Gly	Ala	Ala	Arg	Arg	Pro	Xaa	Ser
			85					90					95		
Gln	Pro	Pro	Leu	His	Arg	His	Arg	Arg	Thr	Ala	Leu	Gln	Pro	Gln	Gly
		100					105					110			
Arg	Trp	Ile	Ala	Arg	Arg	Xaa	Arg	Gly	Val	Arg	Xaa	Arg	Arg	Arg	Val
	115					120						125			
Arg	Val														
	130														

(2) INFORMATION FOR SEQ ID NO:215:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 129 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..129
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481861

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:215:

Ile	Cys	Cys	Pro	Ser	Pro	Arg	Ala	Pro	Cys	Pro	Ser	Ser	Pro	Phe	
1				5				10					15		
Pro	Pro	Gly	Pro	Gly	Val	Ser	Val	Pro	Val	Arg	Xaa	Phe	Ser	Gly	Leu
			20					25					30		
Arg	Ser	Pro	Pro	Pro	Ala	Thr	Ala	Ala	Ser	Gln	Leu	Gly	Lys	Ala	Ile
		35				40						45			
Arg	Pro	Leu	Pro	Ser	Thr	Ala	Leu	Thr	Ala	Arg	Gly	Arg	Ser	Asn	Ala
	50				55					60					
Glu	Ala	Val	Arg	Pro	Ser	Arg	Val	Ile	Val	Met	Pro	Gly	Leu	Ala	Ala
65				70					75					80	
Ala	Glu	Gln	Asp	Ala	Val	Ser	Leu	Val	Arg	Arg	Val	Ala	Xaa	Ala	Leu
			85					90					95		
Asn	Arg	Arg	Phe	Thr	Asp	Ile	Val	Ala	Leu	Leu	Phe	Ser	His	Lys	Gly
		100					105					110			
Ala	Gly	Ser	Leu	Gly	Xaa	Val	Ala	Gly	Phe	Xaa	Ile	Ala	Val	Val	Phe
	115					120						125			
Ala															

(2) INFORMATION FOR SEQ ID NO:216:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 129 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..129

(D) OTHER INFORMATION: / Ceres Seq. ID 1481862

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:216:

```
Met Leu Ser Val Thr Ala Arg Ala Ser Leu Pro Phe Phe Thr Leu Ser
1           5           10           15
Ser Arg Thr Arg Arg Leu Arg Ala Arg Ala Xaa Leu Leu Arg Pro Ala
20           25           30
Val Ser Ser Thr Gly His Arg Cys Leu Ala Ile Gly Gln Gly Asn Gln
35           40           45
Thr Ala Pro Ile Asn Arg Pro Asp Arg Ala Arg Lys Ile Lys Arg Gly
50           55           60
Gly Arg Pro Ser Leu Pro Arg Tyr Arg Asp Ala Arg Pro Arg Arg Ser
65           70           75           80
Arg Ala Gly Arg Arg Leu Ala Gly Ala Ala Arg Arg Pro Xaa Ser Gln
85           90           95
Pro Pro Leu His Arg His Arg Arg Thr Ala Leu Gln Pro Gln Gly Arg
100          105          110
Trp Ile Ala Arg Arg Xaa Arg Gly Val Arg Xaa Arg Arg Val Arg
115          120          125
Val
```

(2) INFORMATION FOR SEQ ID NO:217:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 589 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..589

(D) OTHER INFORMATION: / Ceres Seq. ID 1481863

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:217:

```
agcagccgct cccgactttt accttcctac ctgggtgctgt agcatccgcc gcctcccgca      60
gaacccgaag atggcgctcgt cggcgctogac cctcgaaatc gagggcccgcg acgtgggttaa    120
gatagtgtctg cagttctgca aggagaattc gctgcagcag acgttccaga cgctgcaaaa      180
cgagtgccag gtctccctca acaactgttga cagcatcgac accttcattg ccgacatcaa      240
cgccgggctg tgggatgctg tgcttcccca ggctgcacag ctcaagctgc cacgcaagaa      300
gctcgaggac ctctatgagc agattgtgtt ggagatggct gagctccgtg agcttgacac      360
ggcccgtgcc atcctccgcc agacgcaggc catgggtgtt atgaagcagg agcagcctga      420
rcggtacctc cgccttgagc acctccttgt ccgcacatac tttgaccca atgaggccta      480
ccaagaatcb accaaggaga agcggcgagc actgattgct caagctgttg ctttcagaag      540
tctcagtagt acsgccatct cgtcttatgg cactgattgg tcagggttg
```

(2) INFORMATION FOR SEQ ID NO:218:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 196 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..196

(D) OTHER INFORMATION: / Ceres Seq. ID 1481864

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:218:

```
Ala Ala Ala Pro Asp Phe Tyr Leu Pro Thr Trp Cys Cys Ser Ile Arg
1           5           10           15
Arg Leu Pro Gln Asn Pro Lys Met Ala Ser Ser Ala Ser Thr Leu Glu
20           25           30
Ile Glu Ala Arg Asp Val Val Lys Ile Val Leu Gln Phe Cys Lys Glu
```

35	40	45
Asn Ser Leu Gln Gln Thr Phe Gln Thr Leu Gln Asn Glu Cys Gln Val		
50	55	60
Ser Leu Asn Thr Val Asp Ser Ile Asp Thr Phe Ile Ala Asp Ile Asn		
65	70	75
Ala Gly Arg Trp Asp Ala Val Leu Pro Gln Val Ala Gln Leu Lys Leu		80
	85	90
Pro Arg Lys Lys Leu Glu Asp Leu Tyr Glu Gln Ile Val Leu Glu Met		95
	100	105
Ala Glu Leu Arg Glu Leu Asp Thr Ala Arg Ala Ile Leu Arg Gln Thr		110
	115	120
Gln Val Met Gly Val Met Lys Gln Glu Gln Pro Xaa Arg Tyr Leu Arg		125
130	135	140
Leu Glu His Leu Leu Val Arg Thr Tyr Phe Asp Pro Asn Glu Ala Tyr		
145	150	155
Gln Glu Xaa Thr Lys Glu Lys Arg Arg Ala Leu Ile Ala Gln Ala Val		
	165	170
Ala Phe Arg Ser Leu Ser Ser Xaa Ala Ile Ser Ser Tyr Gly Thr Asp		175
	180	185
Trp Ser Gly Leu		190
195		

(2) INFORMATION FOR SEQ ID NO:219:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 173 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..173

(D) OTHER INFORMATION: / Ceres Seq. ID 1481865

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:219:

Met Ala Ser Ser Ala Ser Thr Leu Glu Ile Glu Ala Arg Asp Val Val		
1	5	10
Lys Ile Val Leu Gln Phe Cys Lys Glu Asn Ser Leu Gln Gln Thr Phe		15
	20	25
Gln Thr Leu Gln Asn Glu Cys Gln Val Ser Leu Asn Thr Val Asp Ser		30
	35	40
Ile Asp Thr Phe Ile Ala Asp Ile Asn Ala Gly Arg Trp Asp Ala Val		45
50	55	60
Leu Pro Gln Val Ala Gln Leu Lys Leu Pro Arg Lys Lys Leu Glu Asp		65
	70	75
Leu Tyr Glu Gln Ile Val Leu Glu Met Ala Glu Leu Arg Glu Leu Asp		80
	85	90
Thr Ala Arg Ala Ile Leu Arg Gln Thr Gln Val Met Gly Val Met Lys		95
	100	105
Gln Glu Gln Pro Xaa Arg Tyr Leu Arg Leu Glu His Leu Leu Val Arg		110
	115	120
Thr Tyr Phe Asp Pro Asn Glu Ala Tyr Gln Glu Xaa Thr Lys Glu Lys		125
130	135	140
Arg Arg Ala Leu Ile Ala Gln Ala Val Ala Phe Arg Ser Leu Ser Ser		145
	150	155
Xaa Ala Ile Ser Ser Tyr Gly Thr Asp Trp Ser Gly Leu		160
	165	170

(2) INFORMATION FOR SEQ ID NO:220:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 554 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: DNA (genomic)
(ix) FEATURE:
(A) NAME/KEY: -
(B) LOCATION: 1..554
(D) OTHER INFORMATION: / Ceres Seq. ID 1481873
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:220:
agaatttgat ttgcaaaaac aaactaagtg gtggcaaaga gcgatccaaa tatgccaaat 60
tatagtcaaa aacaatttgg tcttcaattg cattgatttt gcacttcttg tgttgctttt 120
tgatgtgttg gcataaatca ccaaaaaggg ggagattata aggcaaagt gtcccttgggc 180
catttctaaa atgttttggg gattaagtgc ccaacacgtt tgaataagtt cttatgggcc 240
aaataaagtg agaagtgaat atcaaggcac aatgtatgtt tctagactta gtacatcggt 300
ttttgaaggc taatgtgttt tctctaagtg cttgaaacag tgataaaaga gaagaaaagg 360
attgcaaaag agttggctat gtgcagcaaa ctccagttcg gcttggcaca ccgaactgtc 420
cggtggtgca ccggactgtc cggtgcgcca rgctggtctc cggtgaaatg gccactctcg 480
ggactcaaca acgctatgg ctaaaaatca ccggaccgtc cggtgagtca tctacgacga 540
actcattgct ctgc
(2) INFORMATION FOR SEQ ID NO:221:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 44 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
(A) NAME/KEY: peptide
(B) LOCATION: 1..44
(D) OTHER INFORMATION: / Ceres Seq. ID 1481874
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:221:
Glu Phe Asp Leu Gln Lys Gln Thr Lys Trp Trp Gln Arg Ala Ile Gln
1 5 10 15
Ile Cys Gln Ile Val Lys Asn Asn Leu Val Phe Asn Cys Ile Asp
20 25 30
Phe Ala Leu Leu Val Leu Leu Phe Asp Val Leu Ala
35 40
(2) INFORMATION FOR SEQ ID NO:222:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 41 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
(A) NAME/KEY: peptide
(B) LOCATION: 1..41
(D) OTHER INFORMATION: / Ceres Seq. ID 1481875
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:222:
Met Cys Ser Lys Leu Gln Phe Gly Leu Ala His Arg Thr Val Arg Trp
1 5 10 15
Cys Thr Gly Leu Ser Gly Ala Pro Xaa Trp Ser Pro Val Lys Trp Pro
20 25 30
Leu Ser Gly Leu Asn Asn Ala Tyr Gly
35 40
(2) INFORMATION FOR SEQ ID NO:223:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 478 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single

(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: DNA (genomic)
(ix) FEATURE:
 (A) NAME/KEY: -
 (B) LOCATION: 1..478
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481885
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:223:
gcattgcact mnnngggagcg tgcgcrgtag gagtggatcg gaggagcatg trgaggctaa 60
agattggava aggaggaggg ccatggatgc agacggcgag cgacttccat ggccggcagg 120
tttrggagta cgacccggac gccggcaccg acgaggagcg cacaaggtgg agcagcttcg 180
ccgggcgttc acagagaacc gcttccgaag agggaatcgc aggacctcct aatgcgtatg 240
cagttcgctg gacaaaaata tgtrcatgcv gatctdctg crgccaccaa gatagangag 300
gatggcgacg aggtgccgct gacggaggag aggttgamgg aatcgctgar gcgagcdctg 360
ggtttrcatg ctgctctcca agctgaagat ggccactggc cgcttggtga ttacagtvvg 420
gttatgtacc tcatgccgtt ctggattttc gcaactgcaca tcacaggcac ggtcgtatg
(2) INFORMATION FOR SEQ ID NO:224:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 127 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: peptide
 (ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..127
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481886
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:224:
Ile Ala Xaa Xaa Gly Ala Cys Xaa Val Gly Val Asp Arg Arg Ser Met
1 5 10 15
Xaa Arg Leu Lys Ile Gly Xaa Gly Gly Gly Pro Trp Met Gln Thr Ala
20 25 30
Ser Asp Phe His Gly Arg Gln Val Xaa Glu Tyr Asp Pro Asp Ala Gly
35 40 45
Thr Asp Glu Glu Arg Thr Arg Trp Ser Ser Phe Ala Gly Arg Ser Gln
50 55 60
Arg Thr Ala Ser Glu Glu Gly Ile Ala Gly Pro Pro Asn Ala Tyr Ala
65 70 75 80
Val Arg Trp Thr Lys Ile Cys Xaa Cys Xaa Ser Xaa Cys Xaa His Gln
85 90 95
Asp Arg Xaa Gly Trp Arg Arg Gly Ala Ala Asp Gly Gly Glu Val Xaa
100 105 110
Gly Ile Ala Xaa Ala Ser Xaa Gly Xaa His Gly Cys Ser Pro Ser
115 120 125
(2) INFORMATION FOR SEQ ID NO:225:
 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 112 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear
 (ii) MOLECULE TYPE: peptide
 (ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..112
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481887
 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:225:
Met Xaa Arg Leu Lys Ile Gly Xaa Gly Gly Gly Pro Trp Met Gln Thr
1 5 10 15
Ala Ser Asp Phe His Gly Arg Gln Val Xaa Glu Tyr Asp Pro Asp Ala
20 25 30

Gly	Thr	Asp	Glu	Glu	Arg	Thr	Arg	Trp	Ser	Ser	Phe	Ala	Gly	Arg	Ser
	35						40					45			
Gln	Arg	Thr	Ala	Ser	Glu	Glu	Gly	Ile	Ala	Gly	Pro	Pro	Asn	Ala	Tyr
	50					55					60				
Ala	Val	Arg	Trp	Thr	Lys	Ile	Cys	Xaa	Cys	Xaa	Ser	Xaa	Cys	Xaa	His
65					70					75					80
Gln	Asp	Arg	Xaa	Gly	Trp	Arg	Arg	Gly	Ala	Ala	Asp	Gly	Gly	Glu	Val
				85				90						95	
Xaa	Gly	Ile	Ala	Xaa	Ala	Ser	Xaa	Gly	Xaa	His	Gly	Cys	Ser	Pro	Ser
			100					105						110	

(2) INFORMATION FOR SEQ ID NO:226:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 123 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..123

(D) OTHER INFORMATION: / Ceres Seq. ID 1481888

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:226:

Met	Ala	Gly	Arg	Phe	Xaa	Ser	Thr	Thr	Arg	Thr	Pro	Ala	Pro	Thr	Arg
1				5					10					15	
Ser	Ala	Gln	Gly	Gly	Ala	Ala	Ser	Pro	Gly	Val	His	Arg	Glu	Pro	Leu
		20					25					30			
Pro	Lys	Arg	Glu	Ser	Gln	Asp	Leu	Leu	Met	Arg	Met	Gln	Phe	Ala	Gly
	35				40						45				
Gln	Lys	Tyr	Xaa	His	Xaa	Asp	Xaa	Pro	Xaa	Ala	Thr	Lys	Ile	Xaa	Glu
	50				55					60					
Asp	Gly	Asp	Glu	Val	Pro	Leu	Thr	Glu	Glu	Arg	Leu	Xaa	Glu	Ser	Leu
65				70					75						80
Xaa	Arg	Xaa	Leu	Gly	Xaa	Met	Ala	Ala	Leu	Gln	Ala	Glu	Asp	Gly	His
			85					90						95	
Trp	Pro	Pro	Gly	Asp	Tyr	Ser	Xaa	Val	Met	Tyr	Leu	Met	Pro	Phe	Trp
			100					105						110	
Ile	Phe	Ala	Leu	His	Ile	Thr	Gly	Thr	Val	Asp					
			115					120							

(2) INFORMATION FOR SEQ ID NO:227:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 545 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..545

(D) OTHER INFORMATION: / Ceres Seq. ID 1481893

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:227:

atagccttar	cccggcgcgga	gaagaaatcg	tatcctcgcc	agctcttcac	caacagattc	60
gtctcctcgc	ctccgcccggg	tttcgaccag	aacgccgccc	ccagcccacc	agtaattcct	120
ccgggcactg	gtctccacct	cctctgggat	caccacccaa	gaaaagggtg	cgcggcgcac	180
aggcgaccac	tgagatttta	ttctctatat	aacatttggc	tgtaagtggg	ttataatctc	240
tataactctt	aaataagtgc	aaatatctca	atgtcaagtg	tttcaaattc	tattgctgtg	300
ggtcttccaa	gctatgggct	atatctagag	acaaggtttc	tcacgcagac	ctataggaac	360
ttcgcacaga	aatcctctta	caagtattcc	agaatccgtg	cagtcagggg	aaatgggtggg	420

cgtcgaaggc tgggtgacat aatccgaatc attccagaac tctcaaggga ctatttttaa 480
agtcgatcga ggcgagctct ttttggtggc atctcggtgc ttggcggctt ttacgttgca 540
cagac

(2) INFORMATION FOR SEQ ID NO:228:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 74 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..74
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481894

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:228:

Ile Ala Leu Xaa Arg Arg Glu Lys Lys Ser Tyr Pro Arg Gln Leu Phe
1 5 10 15
Thr Asn Arg Phe Val Ser Ser Pro Pro Pro Gly Phe Asp Gln Asn Ala
20 25 30
Ala Ala Ser Pro Pro Val Ile Pro Pro Gly Thr Gly Leu His Leu Leu
35 40 45
Trp Asp His His Pro Arg Lys Gly Cys Ala Ala His Arg Arg Pro Leu
50 55 60
Arg Phe Tyr Ser Leu Tyr Asn Ile Trp Leu
65 70

(2) INFORMATION FOR SEQ ID NO:229:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 59 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..59
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481895

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:229:

Ser Leu Xaa Pro Ala Arg Glu Glu Ile Val Ser Ser Pro Ala Leu His
1 5 10 15
Gln Gln Ile Arg Leu Leu Ala Ser Ala Gly Phe Arg Pro Glu Arg Arg
20 25 30
Arg Gln Pro Thr Ser Asn Ser Ser Gly His Trp Ser Pro Pro Pro Leu
35 40 45
Gly Ser Pro Pro Lys Lys Arg Val Arg Gly Ala
50 55

(2) INFORMATION FOR SEQ ID NO:230:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 91 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..91
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481896

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:230:

Met Ser Ser Val Ser Asn Ser Ile Ala Val Gly Leu Pro Ser Tyr Gly
1 5 10 15

Leu Tyr Leu Glu Thr Arg Phe Leu Thr Gln Thr Tyr Arg Asn Phe Ala
20 25 30
Gln Lys Ser Ser Tyr Lys Tyr Ser Arg Ile Arg Ala Val Gln Gly Asn
35 40 45
Gly Gly Arg Arg Arg Leu Val Asp Ile Ile Arg Ile Ile Pro Glu Leu
50 55 60
Ser Arg Asp Tyr Phe Lys Ser Arg Ser Arg Arg Ala Leu Phe Gly Gly
65 70 75 80
Ile Ser Leu Leu Gly Gly Phe Tyr Val Ala Gln
85 90

(2) INFORMATION FOR SEQ ID NO:231:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 391 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..391
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481897

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:231:

tatgtggaaa ccatagctat tgggggcgaa gggcttatga gcgctcatttc aaggaatggc	60
gtcatcagca tgggatgcga tgccttgsc a tcccccaata ctaagaattt caatgaaatt	120
acatccatcg aggaggcgaa agcgctctgg gagaaaattc aagcacgaca aggggtgaat	180
aagtggcggc cagacctaga ggaagagtat gaagatcagg aaggcaacat ctacaacaag	240
aagacctaca ctgacctgca gcgtcaaggc ctgatctagg gctcctgctg gttaaagtgtg	300
tcgggatttg ttcagaactt atctcatgta gttgtaactc tgaaaatatt ggcccatctg	360
gcatacatatt tatgtaataa catgattctc c	

(2) INFORMATION FOR SEQ ID NO:232:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 92 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..92
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481898

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:232:

Tyr Val Glu Thr Ile Ala Ile Gly Gly Glu Gly Leu Met Ser Val Ile
1 5 10 15
Ser Arg Asn Gly Val Ile Ser Met Gly Cys Asp Ala Leu Xaa Phe Pro
20 25 30
Asn Thr Lys Asn Phe Asn Glu Ile Thr Ser Ile Glu Glu Ala Lys Ala
35 40 45
Leu Trp Glu Lys Ile Gln Ala Arg Gln Gly Val Asn Lys Trp Arg Pro
50 55 60
Asp Leu Glu Glu Glu Tyr Glu Asp Gln Glu Gly Asn Ile Tyr Asn Lys
65 70 75 80
Lys Thr Tyr Thr Asp Leu Gln Arg Gln Gly Leu Ile
85 90

(2) INFORMATION FOR SEQ ID NO:233:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 80 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..80
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481899

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:233:

```
Met Ser Val Ile Ser Arg Asn Gly Val Ile Ser Met Gly Cys Asp Ala
1          5          10          15
Leu Xaa Phe Pro Asn Thr Lys Asn Phe Asn Glu Ile Thr Ser Ile Glu
          20          25          30
Glu Ala Lys Ala Leu Trp Glu Lys Ile Gln Ala Arg Gln Gly Val Asn
          35          40          45
Lys Trp Arg Pro Asp Leu Glu Glu Tyr Glu Asp Gln Glu Gly Asn
          50          55          60
Ile Tyr Asn Lys Lys Thr Tyr Thr Asp Leu Gln Arg Gln Gly Leu Ile
65          70          75          80
```

(2) INFORMATION FOR SEQ ID NO:234:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 69 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..69
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481900

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:234:

```
Met Gly Cys Asp Ala Leu Xaa Phe Pro Asn Thr Lys Asn Phe Asn Glu
1          5          10          15
Ile Thr Ser Ile Glu Glu Ala Lys Ala Leu Trp Glu Lys Ile Gln Ala
          20          25          30
Arg Gln Gly Val Asn Lys Trp Arg Pro Asp Leu Glu Glu Glu Tyr Glu
          35          40          45
Asp Gln Glu Gly Asn Ile Tyr Asn Lys Lys Thr Tyr Thr Asp Leu Gln
          50          55          60
Arg Gln Gly Leu Ile
65
```

(2) INFORMATION FOR SEQ ID NO:235:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 722 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: DNA (genomic)
(ix) FEATURE:
 (A) NAME/KEY: -
 (B) LOCATION: 1..722
 (D) OTHER INFORMATION: / Ceres Seq. ID 1481901

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:235:

```
aattttattct caaaccttat gagttagatc tctcttaatc attctctctt tcttctctcc      60
tctgtgatgt gaggtttcga agatccttct ctgattcctc atcaaaactca gatcagtagc      120
ggacccaagt cattccttta gagagatata tggcagagggt gaaggatcaa ttagagatta      180
agttccggct taacgatgggt tctgatatcg gtcctaaatt gtttcctgat gctactaccg      240
ttgctacatt gaaagaaacc gttgttgctc agtggccaag agataaggag aacggggccaa      300
agacagtgaa agatgttaaa ctgataagcg cgggtagaat attggagAAC aacaaaacgg      360
ttggagattg caggagtcgc gtcggcaatt tctcagggtgc tgtcaccaca atgcatgtta      420
```

```
taattcaaca tcaagttact gaaaaagaaa agaagaagaa gaagcctaaa ggtgatctga      480
aacagaacaa atgtgtctgt ttatgttttg gagctcgttg ttaacaattg tgcaagacaa      540
gtagagagag ttaaaaaaagc ttgggagatt cacattctgt tcttgagcct tcttcaatac      600
cttttgcctt tgttccttgt aattcttttt tctgacatga catgacatga ttggcttttt      660
gatcgcttga ggtttggttt ctattgtatt tcggattcgc aacaccgtgg aaattattag      720
gt
```

(2) INFORMATION FOR SEQ ID NO:236:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 124 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..124

(D) OTHER INFORMATION: / Ceres Seq. ID 1481902

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:236:

```
Met Ala Glu Val Lys Asp Gln Leu Glu Ile Lys Phe Arg Leu Asn Asp
1          5          10          15
Gly Ser Asp Ile Gly Pro Lys Leu Phe Pro Asp Ala Thr Thr Val Ala
20          25          30
Thr Leu Lys Glu Thr Val Val Ala Gln Trp Pro Arg Asp Lys Glu Asn
35          40          45
Gly Pro Lys Thr Val Lys Asp Val Lys Leu Ile Ser Ala Gly Arg Ile
50          55          60
Leu Glu Asn Asn Lys Thr Val Gly Asp Cys Arg Ser Pro Val Gly Asn
65          70          75          80
Phe Ser Gly Ala Val Thr Thr Met His Val Ile Ile Gln His Gln Val
85          90          95
Thr Glu Lys Glu Lys Lys Lys Lys Lys Pro Lys Gly Asp Leu Lys Gln
100         105         110
Asn Lys Cys Val Cys Leu Cys Phe Gly Ala Arg Cys
115         120
```

(2) INFORMATION FOR SEQ ID NO:237:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 647 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..647

(D) OTHER INFORMATION: / Ceres Seq. ID 1481903

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:237:

```
cccccttta ctacacactt cttctttttt cttcagaaaag aaagaaagac agagagagag      60
agagaagatg gtgttaggaa agcgtcatgg atcactgacg aagagaacaa ctagcatgaa      120
gatgatcaca ctcgatacac ccacgatcta tgacgcatct cagccgtccg atcatctaac      180
ctttcatcaa caccctcaca atccgatggg ggtgatggct agtaactacg atgatttctt      240
gaagacttgk agtctctgca atcgaagtct ctgccatcat cgtgacattt acatgtatag      300
agggaaacaac gcatttttga gcttagaatg caggggagaag caaattaagc tggacgagaa      360
aaaagcgaaa accgggcttcg taacatcgaa gaaaccaatt cgtatttagt tgatcatcta      420
tgatctaaaa tgataacgat agtttttcct tatgagtaaa atgaatatgt tttkcggtwt      480
cgtgtacaag aatgatgaaa ataaagagag aaaaaatgag actaaatgag tgtagtgatc      540
atatagtaat gggacttcac aagcatgatt tgatttggtc gtgtgatttg tttctttgtg      600
atgtgtaata tgtaatgtaa tatcaatggt gatgtatatt caggtggt
```

(2) INFORMATION FOR SEQ ID NO:238:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 135 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..135
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481904
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:238:

```
Pro Pro Leu Leu His Thr Ser Ser Phe Phe Phe Arg Lys Lys Glu Arg
1          5          10          15
Gln Arg Glu Arg Glu Lys Met Val Leu Gly Lys Arg His Gly Ser Leu
20          25          30
Ile Lys Arg Thr Thr Ser Met Lys Met Ile Thr Leu Asp Thr Pro Thr
35          40          45
Ile Tyr Asp Ala Ser Gln Pro Ser Asp His Leu Thr Phe His Gln His
50          55          60
Pro His Asn Pro Met Val Val Met Ala Ser Asn Tyr Asp Asp Phe Leu
65          70          75          80
Lys Thr Xaa Ser Leu Cys Asn Arg Ser Leu Cys His His Arg Asp Ile
85          90          95
Tyr Met Tyr Arg Gly Asn Asn Ala Phe Cys Ser Leu Glu Cys Arg Glu
100         105         110
Lys Gln Ile Lys Leu Asp Glu Lys Lys Ala Lys Thr Gly Phe Val Thr
115         120         125
Ser Lys Lys Pro Ile Arg Ile
130         135
```

(2) INFORMATION FOR SEQ ID NO:239:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 113 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..113
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481905
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:239:

```
Met Val Leu Gly Lys Arg His Gly Ser Leu Ile Lys Arg Thr Thr Ser
1          5          10          15
Met Lys Met Ile Thr Leu Asp Thr Pro Thr Ile Tyr Asp Ala Ser Gln
20          25          30
Pro Ser Asp His Leu Thr Phe His Gln His Pro His Asn Pro Met Val
35          40          45
Val Met Ala Ser Asn Tyr Asp Asp Phe Leu Lys Thr Xaa Ser Leu Cys
50          55          60
Asn Arg Ser Leu Cys His His Arg Asp Ile Tyr Met Tyr Arg Gly Asn
65          70          75          80
Asn Ala Phe Cys Ser Leu Glu Cys Arg Glu Lys Gln Ile Lys Leu Asp
85          90          95
Glu Lys Lys Ala Lys Thr Gly Phe Val Thr Ser Lys Lys Pro Ile Arg
100         105         110
Ile
```

(2) INFORMATION FOR SEQ ID NO:240:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 97 amino acids

- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..97
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481906

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:240:

```
Met Lys Met Ile Thr Leu Asp Thr Pro Thr Ile Tyr Asp Ala Ser Gln
 1             5             10             15
Pro Ser Asp His Leu Thr Phe His Gln His Pro His Asn Pro Met Val
          20          25          30
Val Met Ala Ser Asn Tyr Asp Asp Phe Leu Lys Thr Xaa Ser Leu Cys
          35          40          45
Asn Arg Ser Leu Cys His His Arg Asp Ile Tyr Met Tyr Arg Gly Asn
          50          55          60
Asn Ala Phe Cys Ser Leu Glu Cys Arg Glu Lys Gln Ile Lys Leu Asp
          65          70          75          80
Glu Lys Lys Ala Lys Thr Gly Phe Val Thr Ser Lys Lys Pro Ile Arg
          85          90          95
Ile
```

(2) INFORMATION FOR SEQ ID NO:241:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 800 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- (ix) FEATURE:
 - (A) NAME/KEY: -
 - (B) LOCATION: 1..800
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481907

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:241:

```
actttcatta gtttccaatt taacaaatca aaatcagaag aagaagaaga tgaccagctc      60
tgatcctcaa tctcacaacg tcttcgtcta cggtagcatt ctagaacccg ccgtcgccgc      120
cgtgatcctt gatcgcacag ccgatacagt ccccgccggt ctccatggct agtacgctct      180
ctcacctctt cgatgatcgt ttattcaatc ggagattaac aaaagattta tgggttttta      240
acagtcacag atataaaatc aaaggacttc catatccatg tattgtttct tctgattctg      300
gaaaagtcaa cggaaagggt ataactggag tgtctgatgc tgagttaaat aatttcgatg      360
tgattgaagg taatgattat gagagagtaa ctggtgaagt tgtaagaatg gataattctg      420
agaaggtgaa agttgaaact tatgtttggg ttaataaaga tgatcctaga atgtatggag      480
aatgggattt cgaggaatgg agagtgggtc acgcggagaa attcgtggag acttttagaa      540
aaatgttgga atggaacaag aatccaaatg ggaagagcat ggaggaggct gtaggatcat      600
tattatcgtc aggggattaa ttcttgatga gcttggctaa tcttagcaga agagagtaag      660
tgagtaagta aagagtgggt tttgaataat gtgttggttg aacttgaaac agagtcttct      720
atgtgatttt gtttgtgttg ttatggatat cttgttggca ctttttctga tttcagttgg      780
aaacaggtgc gtttgcgggc
```

(2) INFORMATION FOR SEQ ID NO:242:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 151 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..151

(D) OTHER INFORMATION: / Ceres Seq. ID 1481908

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:242:

```
Met Ala Ser Thr Leu Ser His Pro Leu Asp Asp Arg Leu Phe Asn Arg
1           5           10           15
Arg Leu Thr Lys Asp Leu Trp Val Phe Asn Ser His Arg Tyr Lys Leu
          20           25           30
Lys Gly Leu Pro Tyr Pro Cys Ile Val Ser Ser Asp Ser Gly Lys Val
          35           40           45
Asn Gly Lys Val Ile Thr Gly Val Ser Asp Ala Glu Leu Asn Asn Phe
          50           55           60
Asp Val Ile Glu Gly Asn Asp Tyr Glu Arg Val Thr Val Glu Val Val
          65           70           75           80
Arg Met Asp Asn Ser Glu Lys Val Lys Val Glu Thr Tyr Val Trp Val
          85           90           95
Asn Lys Asp Asp Pro Arg Met Tyr Gly Glu Trp Asp Phe Glu Glu Trp
          100          105          110
Arg Val Val His Ala Glu Lys Phe Val Glu Thr Phe Arg Lys Met Leu
          115          120          125
Glu Trp Asn Lys Asn Pro Asn Gly Lys Ser Met Glu Glu Ala Val Gly
          130          135          140
Ser Leu Leu Ser Ser Gly Asp
          145          150
```

(2) INFORMATION FOR SEQ ID NO:243:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 675 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..675

(D) OTHER INFORMATION: / Ceres Seq. ID 1481913

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:243:

```
aactcagtcg tctcaatgcc gtgtaacaac ttcacgtct ttttcagttc aacaacttca      60
tcgtcttttt cagttcaacc tccgacatct ctcgtctcca ggtgattgat cccatcgaag      120
ggtctatgga agagaacaac aacgccggga gcgattccga ctctaattcc gtcgaagatt      180
cacaagacta ttacgaaccg atctcagccg tcgatttata taactccaac gacgatgaag      240
aagacagtta tcttccgatac ggtggagatg gtctctctaa cggccattgt atgattccgg      300
atgcagaggt aggaatctct tctattagta taaacgataa cacagacagc gaagaagaga      360
cagagacgga gactggaccg gagatccgta gagcgtttga ggaggacgaa cggcggagaa      420
ggtcgccgtt agtggaggag aatgccgtta gggttatgga ggcaatgcga gccatctcat      480
tccctggaac ggctcctgat tgggcctccg atgttaatga ggatcgttgg attgatcagc      540
tgccaagatt gagaaccact tctcaataag ctttctccaa tctcgatagt tgttttcgtt      600
taagatcttt ctcaatgttg ttcaatgtga cttcttttaa acattcaata taaaaaccag      660
agaatttcac cactc
```

(2) INFORMATION FOR SEQ ID NO:244:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 147 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..147

(D) OTHER INFORMATION: / Ceres Seq. ID 1481914

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:244:

```
Met Glu Glu Asn Asn Asn Ala Gly Ser Asp Ser Asp Ser Asn Ser Val
```



```

1           5           10           15
Glu Asp Ser Gln Asp Tyr Tyr Glu Pro Ile Ser Ala Val Asp Leu Tyr
                20           25           30
Asn Ser Asn Asp Asp Glu Glu Asp Ser Tyr Leu Pro Ile Gly Gly Asp
                35           40           45
Gly Leu Ser Asn Gly His Cys Met Ile Pro Asp Ala Glu Val Gly Ile
                50           55           60
Ser Ser Ile Ser Ile Asn Asp Asn Thr Asp Ser Glu Glu Glu Thr Glu
65                70           75           80
Thr Glu Thr Gly Pro Glu Ile Arg Arg Ala Phe Glu Glu Asp Glu Arg
                85           90           95
Arg Arg Arg Ser Pro Leu Val Glu Glu Asn Ala Val Arg Val Met Glu
                100          105          110
Ala Met Arg Ala Ile Ser Phe Pro Gly Thr Ala Pro Asp Trp Ala Ser
                115          120          125
Asp Val Asn Glu Asp Arg Trp Ile Asp Gln Leu Arg Arg Leu Arg Thr
130                135          140
Thr Ser Gln
145
```

(2) INFORMATION FOR SEQ ID NO:245:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 92 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..92
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481915

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:245:

```

Met Ile Pro Asp Ala Glu Val Gly Ile Ser Ser Ile Ser Ile Asn Asp
1           5           10           15
Asn Thr Asp Ser Glu Glu Glu Thr Glu Thr Glu Thr Gly Pro Glu Ile
                20           25           30
Arg Arg Ala Phe Glu Glu Asp Glu Arg Arg Arg Arg Ser Pro Leu Val
                35           40           45
Glu Glu Asn Ala Val Arg Val Met Glu Ala Met Arg Ala Ile Ser Phe
50                55           60
Pro Gly Thr Ala Pro Asp Trp Ala Ser Asp Val Asn Glu Asp Arg Trp
65                70           75           80
Ile Asp Gln Leu Arg Arg Leu Arg Thr Thr Ser Gln
                85           90
```

(2) INFORMATION FOR SEQ ID NO:246:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 729 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..729
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481916

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:246:

```

acgattttta tctgatttga caccaaagta tcttttagcc ttaattcggt acgttgtaaa      60
gaaactgatc caatcccttc tattcggatt atatagacct aatatgtaca gatccgcgag      120
ctggaaccgt gtgacggagg attactcggg gccttggtcc gcaccaaagg gattatggaa      180
gggcttagac gaagacgagg ccggctccat acgatcccac tggccaaaag atgactaaga      240
```

```
aagagaagtc acgtaccaag tttgctgaaa acgccgttca cataatccct tttgtccttc 300
ttgcttgtgc tctcgtcctt tggttcttct ctaatccaga tgtagatgtt ggggtgaaag 360
gggacttcat tgcggctagg attgaaggat taacgatcga aggagacatt gacaatgaca 420
gcgacggamc tcagaccgga ttcttaggag ccgccacaga ggtcggacat tcaaaaaata 480
aactaaaacg cgaggctaataaacgcaatc ggaggatata agcttcaagg aaagtgatga 540
aaggttttta ttaatcacct ttttgtttga taaatgttta cgagataaac tttcaaaacg 600
aattattctt ttttttcttt ctattttgat tgcgcagtgt agttgatcag gagatgtgtt 660
tctttggtta aacttttata tttagtctct cacattatct tcaagatoca caagaactac 720
tttactct
```

(2) INFORMATION FOR SEQ ID NO:247:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 78 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..78
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481917

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:247:

```
Arg Phe Leu Ser Asp Leu Thr Pro Lys Tyr Leu Leu Ala Leu Ile Arg
1          5          10          15
Tyr Val Val Lys Lys Leu Ile Gln Ser Leu Leu Phe Gly Leu Tyr Arg
          20          25          30
Pro Asn Met Tyr Arg Ser Ala Ser Trp Asn Arg Val Thr Glu Asp Tyr
          35          40          45
Ser Val Pro Trp Ser Ala Pro Lys Gly Leu Trp Lys Gly Leu Asp Glu
          50          55          60
Asp Glu Ala Gly Ser Ile Arg Ser His Trp Pro Lys Asp Asp
65          70          75
```

(2) INFORMATION FOR SEQ ID NO:248:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 107 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..107
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481918

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:248:

```
Met Thr Lys Lys Glu Lys Ser Arg Thr Lys Phe Ala Glu Asn Ala Val
1          5          10          15
His Ile Ile Pro Phe Val Leu Leu Ala Cys Ala Leu Val Leu Trp Phe
          20          25          30
Phe Ser Asn Pro Asp Val Asp Val Gly Val Lys Gly Asp Phe Ile Ala
          35          40          45
Ala Arg Ile Glu Gly Leu Thr Ile Glu Gly Asp Ile Asp Asn Asp Ser
          50          55          60
Asp Gly Xaa Gln Thr Gly Phe Leu Gly Ala Ala Thr Glu Val Gly His
65          70          75          80
Ser Lys Asn Lys Leu Lys Arg Glu Ala Asn Lys Arg Asn Arg Arg Ile
          85          90          95
Gln Ala Ser Arg Lys Val Met Lys Gly Phe Tyr
          100          105
```

(2) INFORMATION FOR SEQ ID NO:249:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 674 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- (ix) FEATURE:
 - (A) NAME/KEY: -
 - (B) LOCATION: 1..674
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481919

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:249:

```
ctgaacgaag ctctctctct gattggccgg atctgccgga gagaaaaatg acgacgagta      60
ttcacatcac agctctcgac ggaatcgta acgtgaactc actcttcaca ctgcgcgtat      120
tcacatcgatt agcttggaac cctaccgatc cagacaacag cctcgtaacc gaccctaatt      180
gcgtcccccac agctcgtatg gctgagaatc tcgtcgccctt ccatgtgtac tctttcgcac      240
cattcctatt ctcaagtctc atcgctctag gtctcaaaaca agcaatgagg ctcaacatag      300
cttcttcggtt tcacatctct actcgaatcg atcctgtggt ttactatgtg aacaagacgg      360
ctcttagatt tgggatgggt acatccgggt tgggatcggt ttgtggatgt gggtttctca      420
tgttggccttt gattaatggt gttcagatca agcttgggac tttgggctgt ggtgctagtg      480
gtcataactta tgcagctggt gtgccgcttt gtgattctgg ttccttctgc acttttcatc      540
tatgtttctc ttatgttata tgcttttact cgttagagac atggtttttg attccatggt      600
tgatgcaatt agggttatgt ttgtatgatg atgatatgat ggaaatgaga atgattctat      660
gytttgatat gggt
```

(2) INFORMATION FOR SEQ ID NO:250:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 224 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..224
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481920

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:250:

```
Glu Arg Ser Ser Leu Ser Asp Trp Pro Asp Leu Pro Glu Arg Lys Met
1          5          10          15
Thr Thr Ser Ile His Ile Thr Ala Leu Asp Gly Ile Val Asn Val Asn
20        25        30
Ser Leu Phe Thr Leu Ala Val Phe Ile Gly Leu Ala Trp Asn Pro Thr
35        40        45
Asp Pro Asp Asn Ser Leu Val Thr Asp Pro Asn Cys Val Pro Thr Ala
50        55        60
Arg Met Ala Glu Asn Leu Val Ala Phe His Val Tyr Ser Phe Ala Ser
65        70        75        80
Phe Leu Phe Ser Ser Leu Ile Ala Leu Gly Leu Lys Gln Ala Met Arg
85        90        95
Leu Asn Ile Ala Ser Ser Phe His Ile Ser Thr Arg Ile Asp Pro Val
100       105       110
Val Tyr Tyr Val Asn Lys Thr Ala Leu Arg Phe Gly Met Val Thr Ser
115       120       125
Gly Leu Gly Ser Val Cys Gly Cys Gly Phe Leu Met Leu Ala Leu Ile
130       135       140
Asn Val Val Gln Ile Lys Leu Gly Thr Leu Gly Cys Gly Ala Ser Gly
145       150       155       160
His Thr Tyr Ala Ala Val Val Pro Leu Cys Asp Ser Gly Ser Phe Cys
165       170       175
Thr Phe His Leu Cys Phe Ser Tyr Val Ile Cys Phe Tyr Ser Leu Glu
180       185       190
Thr Trp Phe Leu Ile Pro Trp Leu Met Gln Leu Gly Leu Cys Leu Tyr
```

	195		200		205										
Asp	Asp	Asp	Met	Met	Glu	Met	Arg	Met	Ile	Leu	Xaa	Phe	Asp	Met	Val
	210				215					220					

(2) INFORMATION FOR SEQ ID NO:251:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 209 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..209
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481921

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:251:

Met	Thr	Thr	Ser	Ile	His	Ile	Thr	Ala	Leu	Asp	Gly	Ile	Val	Asn	Val
1				5					10					15	
Asn	Ser	Leu	Phe	Thr	Leu	Ala	Val	Phe	Ile	Gly	Leu	Ala	Trp	Asn	Pro
			20					25					30		
Thr	Asp	Pro	Asp	Asn	Ser	Leu	Val	Thr	Asp	Pro	Asn	Cys	Val	Pro	Thr
		35					40					45			
Ala	Arg	Met	Ala	Glu	Asn	Leu	Val	Ala	Phe	His	Val	Tyr	Ser	Phe	Ala
	50					55					60				
Ser	Phe	Leu	Phe	Ser	Ser	Leu	Ile	Ala	Leu	Gly	Leu	Lys	Gln	Ala	Met
65					70				75					80	
Arg	Leu	Asn	Ile	Ala	Ser	Ser	Phe	His	Ile	Ser	Thr	Arg	Ile	Asp	Pro
			85					90						95	
Val	Val	Tyr	Tyr	Val	Asn	Lys	Thr	Ala	Leu	Arg	Phe	Gly	Met	Val	Thr
			100					105					110		
Ser	Gly	Leu	Gly	Ser	Val	Cys	Gly	Cys	Gly	Phe	Leu	Met	Leu	Ala	Leu
	115					120						125			
Ile	Asn	Val	Val	Gln	Ile	Lys	Leu	Gly	Thr	Leu	Gly	Cys	Gly	Ala	Ser
	130					135					140				
Gly	His	Thr	Tyr	Ala	Ala	Val	Val	Pro	Leu	Cys	Asp	Ser	Gly	Ser	Phe
145					150					155				160	
Cys	Thr	Phe	His	Leu	Cys	Phe	Ser	Tyr	Val	Ile	Cys	Phe	Tyr	Ser	Leu
			165					170						175	
Glu	Thr	Trp	Phe	Leu	Ile	Pro	Trp	Leu	Met	Gln	Leu	Gly	Leu	Cys	Leu
		180						185					190		
Tyr	Asp	Asp	Asp	Met	Met	Glu	Met	Arg	Met	Ile	Leu	Xaa	Phe	Asp	Met
		195				200						205			
Val															

(2) INFORMATION FOR SEQ ID NO:252:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 159 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..159
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481922

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:252:

Met	Ala	Glu	Asn	Leu	Val	Ala	Phe	His	Val	Tyr	Ser	Phe	Ala	Ser	Phe
1				5						10				15	

Leu Phe Ser Ser Leu Ile Ala Leu Gly Leu Lys Gln Ala Met Arg Leu
20 25 30
Asn Ile Ala Ser Ser Phe His Ile Ser Thr Arg Ile Asp Pro Val Val
35 40 45
Tyr Tyr Val Asn Lys Thr Ala Leu Arg Phe Gly Met Val Thr Ser Gly
50 55 60
Leu Gly Ser Val Cys Gly Cys Gly Phe Leu Met Leu Ala Leu Ile Asn
65 70 75 80
Val Val Gln Ile Lys Leu Gly Thr Leu Gly Cys Gly Ala Ser Gly His
85 90 95
Thr Tyr Ala Ala Val Val Pro Leu Cys Asp Ser Gly Ser Phe Cys Thr
100 105 110
Phe His Leu Cys Phe Ser Tyr Val Ile Cys Phe Tyr Ser Leu Glu Thr
115 120 125
Trp Phe Leu Ile Pro Trp Leu Met Gln Leu Gly Leu Cys Leu Tyr Asp
130 135 140
Asp Asp Met Met Glu Met Arg Met Ile Leu Xaa Phe Asp Met Val
145 150 155

(2) INFORMATION FOR SEQ ID NO:253:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 724 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..724
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481923

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:253:

aaaaagaagg atcaagaacc caaaatagag agcccaattt ctcttaaact tgccaaagta	60
gctatcaggt ggttcttgat acggaacttc cagatcccaa gcagcagcag aggcctcatc	120
accgtcgcct ccaccgagga aatctccggc agcgattctt gacttgatga aacggagctt	180
gtgagtggcg agaccgagtg agcttatggc agcgacgctg gtacttgaag ccgttgatgc	240
aacggagagt ttgagcgacg gagacgtaga agaagaggtg agagaaggat aggaggcaga	300
tgggtgccca gcaatggcgg cagtgaagac gacctgaggg ttgagagagg gagaagaaga	360
ttacggcgag gaagatgaag aagagctgaa atagcttggt ggagcttctt cttctggtgg	420
tcaatggctc gtttcttctc taagggtttc tcgaagtggg gctggattat tgagtttagt	480
gctttagtag cagtttcttt ggggatgaaa ggtttggttat tctgggtcaat ttcgtcgtcg	540
tagtccgcca ttgaaggact gagaagagag aaaaagtgtt attggttaga gagatggttt	600
ggggattgtg tgtagygaac atgtgggtgt ggtgtcgtat ctctagacaa gtattatcca	660
tctcaacggt cttgttctga tttttgatgt tttgtccgta ctcaataaat attttactgg	720

gagt

(2) INFORMATION FOR SEQ ID NO:254:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 53 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..53
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481924

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:254:

Lys Lys Lys Asp Gln Glu Pro Lys Ile Glu Ser Pro Ile Ser Leu Lys
1 5 10 15
Leu Ala Lys Val Ala Ile Arg Trp Phe Leu Ile Arg Asn Phe Gln Ile
20 25 30

Pro Ser Ser Ser Arg Gly Leu Ile Thr Val Ala Ser Thr Glu Glu Ile
35 40 45
Ser Gly Ser Asp Ser
50

(2) INFORMATION FOR SEQ ID NO:255:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 39 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..39
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481925

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:255:

Met Ala Arg Phe Phe Ser Lys Val Phe Ser Lys Trp Ser Trp Ile Ile
1 5 10 15
Glu Phe Ser Ala Cys Arg Ala Val Ser Leu Gly Met Lys Gly Trp Leu
20 25 30
Phe Trp Ser Ile Ser Ser Ser
35

(2) INFORMATION FOR SEQ ID NO:256:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 453 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..453
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481941

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:256:

gcacgcacatc gatcctccca tctgcgcacc cgcaagcyta ttgcgcgcac ctccctcaggt 60
gaccgggaag atgatgccgt tgagccaaac cgacttctcg cgtgcgcagt tcacctcctc 120
ccagaatgcc gccgccgact ccaccacgcc ttccaagatg cgcggcgcggt ccagcaccat 180
gcctctcacc gtgaagcagg tcgtcgacgc gcacgagtct ggcacgggcg acaagggcgc 240
tccgttcacg gtcaatggcg tcgagatggc taacgtaccg ataatcctct tgttcgtcct 300
ttggtccggt gatatgcaga tgttctcggc gttaattcat ctgccgcggg tcccttttca 360
gattcgactt gtggggatgg tcaatgccaa ggtggagcgg acgaccgatg tgaccttcac 420
gctcgacgat ggcaccggcc gcctcgattt cat

(2) INFORMATION FOR SEQ ID NO:257:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 150 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..150
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481942

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:257:

His Arg Ile Asp Pro Pro Ile Cys Ala Pro Ala Ser Xaa Phe Ala Ala
1 5 10 15
Pro Pro Gln Val Thr Gly Lys Met Met Pro Leu Ser Gln Thr Asp Phe
20 25 30
Ser Pro Ser Gln Phe Thr Ser Ser Gln Asn Ala Ala Ala Asp Ser Thr

```

      35              40              45
Thr Pro Ser Lys Met Arg Gly Ala Ser Ser Thr Met Pro Leu Thr Val
  50              55              60
Lys Gln Val Val Asp Ala His Glu Ser Gly Thr Gly Asp Lys Gly Ala
  65              70              75              80
Pro Phe Ile Val Asn Gly Val Glu Met Ala Asn Val Pro Ile Ile Leu
      85              90              95
Leu Phe Val Leu Trp Ser Val Asp Met Gln Met Phe Ser Ala Leu Ile
      100              105              110
His Leu Pro Arg Phe Pro Phe Gln Ile Arg Leu Val Gly Met Val Asn
      115              120              125
Ala Lys Val Glu Arg Thr Thr Asp Val Thr Phe Thr Leu Asp Asp Gly
      130              135              140
Thr Gly Arg Leu Asp Phe
  145              150
```

(2) INFORMATION FOR SEQ ID NO:258:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 127 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..127
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481943

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:258:

```

Met Met Pro Leu Ser Gln Thr Asp Phe Ser Pro Ser Gln Phe Thr Ser
  1              5              10              15
Ser Gln Asn Ala Ala Ala Asp Ser Thr Thr Pro Ser Lys Met Arg Gly
      20              25              30
Ala Ser Ser Thr Met Pro Leu Thr Val Lys Gln Val Val Asp Ala His
      35              40              45
Glu Ser Gly Thr Gly Asp Lys Gly Ala Pro Phe Ile Val Asn Gly Val
      50              55              60
Glu Met Ala Asn Val Pro Ile Ile Leu Leu Phe Val Leu Trp Ser Val
      65              70              75              80
Asp Met Gln Met Phe Ser Ala Leu Ile His Leu Pro Arg Phe Pro Phe
      85              90              95
Gln Ile Arg Leu Val Gly Met Val Asn Ala Lys Val Glu Arg Thr Thr
      100              105              110
Asp Val Thr Phe Thr Leu Asp Asp Gly Thr Gly Arg Leu Asp Phe
      115              120              125
```

(2) INFORMATION FOR SEQ ID NO:259:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 126 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..126
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481944

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:259:

```

Met Pro Leu Ser Gln Thr Asp Phe Ser Pro Ser Gln Phe Thr Ser Ser
  1              5              10              15
Gln Asn Ala Ala Ala Asp Ser Thr Thr Pro Ser Lys Met Arg Gly Ala
      20              25              30
```

Ser Ser Thr Met Pro Leu Thr Val Lys Gln Val Val Asp Ala His Glu
35 40 45
Ser Gly Thr Gly Asp Lys Gly Ala Pro Phe Ile Val Asn Gly Val Glu
50 55 60
Met Ala Asn Val Pro Ile Ile Leu Leu Phe Val Leu Trp Ser Val Asp
65 70 75 80
Met Gln Met Phe Ser Ala Leu Ile His Leu Pro Arg Phe Pro Phe Gln
85 90 95
Ile Arg Leu Val Gly Met Val Asn Ala Lys Val Glu Arg Thr Thr Asp
100 105 110
Val Thr Phe Thr Leu Asp Asp Gly Thr Gly Arg Leu Asp Phe
115 120 125

(2) INFORMATION FOR SEQ ID NO:260:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 677 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..677
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481949

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:260:

acattctagt	acaatatagt	ggttgtgctc	ctctattcta	tttccttggt	gctactagtc	60
tgagttgtga	gattagtgtt	gctaacaatt	tggaagacgc	ggastocctt	tcacctctag	120
caaggttctc	caaatcgctc	gctaaatttt	acaggcgctc	ccagagccgc	taattgtcgt	180
ggatcttcag	acgtccgcta	cacgccgatt	cactccctct	cccgcgctag	ggcggaacct	240
tctcccttgc	gtcttcccat	cgcaaggtct	tgtccatgcc	gacagctagt	tcccgacgga	300
cttcctcgga	ggcggtcagc	accgacgacg	aggaggctgc	gcggggaagc	aagggcggac	360
gaccctcgcc	gccgcgctgc	tcgtcgtgca	ggtagtcggc	tacttcttac	acggtcgccg	420
ccggtgttgg	gctctccgac	agtgcgtgca	tcgatggtgc	agactctctg	cacagccacg	480
ccgatgagct	ctcctctggt	gtcgtggaca	tgcttcacgg	ttcctcscct	gcggccacaa	540
caagcgatgg	tggtggctgg	tgcgctctag	gtgctcgatg	aaaggtgtgt	ttgtagttcg	600
gcacttttta	ccacaggaaa	gagagagaag	taaacaatat	gcatgcgaag	tcaataaaag	660
tgaaatcgaa	attctttt					

(2) INFORMATION FOR SEQ ID NO:261:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 56 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..56
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481950

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:261:

Ile Leu Val Gln Tyr Ser Gly Cys Ala Pro Leu Phe Tyr Phe Leu Val
1 5 10 15
Ala Thr Ser Leu Ser Cys Glu Ile Ser Val Ala Asn Asn Leu Glu Asp
20 25 30
Ala Xaa Ser Phe Ser Pro Leu Ala Arg Phe Ser Lys Ser Ser Ala Lys
35 40 45
Phe Tyr Arg Arg Arg Gln Ser Arg
50 55

(2) INFORMATION FOR SEQ ID NO:262:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 39 amino acids

(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
(A) NAME/KEY: peptide
(B) LOCATION: 1..39
(D) OTHER INFORMATION: / Ceres Seq. ID 1481951
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:262:
Met Pro Thr Ala Ser Ser Arg Arg Thr Ser Ser Glu Ala Val Ser Thr
1 5 10 15
Asp Asp Glu Glu Ala Ala Arg Gly Ser Lys Gly Gly Arg Pro Ser Pro
20 25 30
Pro Arg Cys Ser Ser Cys Arg
35

(2) INFORMATION FOR SEQ ID NO:263:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 38 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
(A) NAME/KEY: peptide
(B) LOCATION: 1..38
(D) OTHER INFORMATION: / Ceres Seq. ID 1481952
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:263:
Met Val Gln Thr Leu Cys Thr Ala Thr Pro Met Ser Ser Pro Leu Leu
1 5 10 15
Ser Trp Thr Cys Phe Thr Val Pro Xaa Ser Arg Pro Gln Gln Ala Met
20 25 30
Val Val Ala Gly Ala Leu
35

(2) INFORMATION FOR SEQ ID NO:264:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 588 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: DNA (genomic)
(ix) FEATURE:
(A) NAME/KEY: -
(B) LOCATION: 1..588
(D) OTHER INFORMATION: / Ceres Seq. ID 1481965
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:264:
caacttcttg ccattgattc agcagctgca gtgcagctac ttcggagggtc tctgattggt 60
gatgaattaa caggaaaaga aaagaaagcc ctgcgcagaa ccatgactga cctggcgtca 120
gttattccca tcggtattct aatgcttctt cctgttacag cggttgggtca cgctgccatg 180
ctggctggaa ttcagagata tgtaccaggc ctgattcctt ccacatacgg gtccgaaagg 240
ttgaacctat tgagacagct tgagaaaatc aaggaactgc aaacaaatga aaccgagagc 300
gaagaaggcg tagaggaaat agcattatga gtagaaggaa gcaatataga cttgtacctc 360
tattcacttt gttcggtaat tcattgccaa aagctgcgca tagagaatct cgttccatgt 420
gtccgggtact ccgggtaagc accagtgtact gcagtcctga ggagcatctt caggagttcc 480
cggttctcga taccgcgaag gatgagcatc tttcctgaac tcggtcagat atgtaatgtt 540
cagaaactta accttactat gttcatactc catttcttgg attgtctt

(2) INFORMATION FOR SEQ ID NO:265:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 109 amino acids
(B) TYPE: amino acid

- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..109
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481966
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:265:

Gln	Leu	Leu	Ala	Ile	Asp	Ser	Ala	Ala	Ala	Val	Gln	Leu	Leu	Arg	Arg
1			5						10					15	
Ser	Leu	Ile	Gly	Asp	Glu	Leu	Thr	Gly	Lys	Glu	Lys	Lys	Ala	Leu	Arg
			20					25					30		
Arg	Thr	Met	Thr	Asp	Leu	Ala	Ser	Val	Ile	Pro	Ile	Gly	Ile	Leu	Met
		35					40					45			
Leu	Leu	Pro	Val	Thr	Ala	Val	Gly	His	Ala	Ala	Met	Leu	Ala	Gly	Ile
	50					55					60				
Gln	Arg	Tyr	Val	Pro	Gly	Leu	Ile	Pro	Ser	Thr	Tyr	Gly	Ser	Glu	Arg
65					70				75					80	
Leu	Asn	Leu	Leu	Arg	Gln	Leu	Glu	Lys	Ile	Lys	Glu	Leu	Gln	Thr	Asn
				85				90						95	
Glu	Thr	Glu	Ser	Glu	Glu	Gly	Val	Glu	Glu	Ile	Ala	Leu			
				100				105							

- (2) INFORMATION FOR SEQ ID NO:266:
 - (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 75 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
 - (ii) MOLECULE TYPE: peptide
 - (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..75
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481967
 - (xi) SEQUENCE DESCRIPTION: SEQ ID NO:266:

Met	Thr	Asp	Leu	Ala	Ser	Val	Ile	Pro	Ile	Gly	Ile	Leu	Met	Leu	Leu
1			5						10					15	
Pro	Val	Thr	Ala	Val	Gly	His	Ala	Ala	Met	Leu	Ala	Gly	Ile	Gln	Arg
			20					25				30			
Tyr	Val	Pro	Gly	Leu	Ile	Pro	Ser	Thr	Tyr	Gly	Ser	Glu	Arg	Leu	Asn
		35					40				45				
Leu	Leu	Arg	Gln	Leu	Glu	Lys	Ile	Lys	Glu	Leu	Gln	Thr	Asn	Glu	Thr
	50					55					60				
Glu	Ser	Glu	Glu	Gly	Val	Glu	Glu	Ile	Ala	Leu					
65				70				75							

- (2) INFORMATION FOR SEQ ID NO:267:
 - (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 63 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
 - (ii) MOLECULE TYPE: peptide
 - (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..63
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481968
 - (xi) SEQUENCE DESCRIPTION: SEQ ID NO:267:

Met	Ser	Arg	Arg	Lys	Gln	Tyr	Arg	Leu	Val	Pro	Leu	Phe	Thr	Leu	Phe
1				5					10					15	
Gly	Asn	Ser	Leu	Pro	Lys	Ala	Ala	His	Arg	Glu	Ser	Arg	Ser	Met	Cys

20 25 30
Pro Val Leu Arg Val Ser Thr Ser Asp Cys Ser Leu Glu Glu His Leu
35 40 45
Gln Glu Phe Pro Val Leu Asp Thr Ala Lys Asp Glu His Leu Ser
50 55 60

(2) INFORMATION FOR SEQ ID NO:268:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 498 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..498
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481973

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:268:

accatcacga atcgcgattt ttttttgaga ttacggaagc ttcgcttgat ttgggatttt	60
taggggttttc tttctccgaa gacgactccg agagaccaac agtgatttga caatgacgct	120
acctccaggt ctttactccg gcaccagctc tcttgctctg gtggctcgtg cttcggcttt	180
tgggttgggt ctcgtctatg ggaacatgaa gctcaagatc aaatcgatgt cacagaagaa	240
ggttgaagcc accgctcatc attaaaccac tcgttctttc tttacaataa gatgccaaaa	300
gctgggggtg atgtctcccc ggtagttttg atttcttctt tcatgattca tccttttagca	360
taagaaggaa caaatgtgtt ttgaaaagc atattatacg gttttaagac ctttttgag	420
ccataattgc cattggctta aaaccgagt caagaacatc tttccatttg ttgtcatcca	480
ataacaccgt tcacattc	

(2) INFORMATION FOR SEQ ID NO:269:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 35 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..35
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481974

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:269:

His His Glu Ser Arg Phe Phe Phe Glu Ile Thr Glu Ala Ser Leu Asp	
1 5 10 15	
Leu Gly Phe Leu Gly Phe Ser Phe Ser Glu Asp Asp Ser Glu Arg Pro	
20 25 30	
Thr Val Ile	
35	

(2) INFORMATION FOR SEQ ID NO:270:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 50 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..50
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481975

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:270:

Met Thr Leu Pro Pro Gly Leu Tyr Ser Gly Thr Ser Ser Leu Ala Leu	
1 5 10 15	
Val Ala Arg Ala Ser Ala Phe Gly Leu Gly Leu Val Tyr Gly Asn Met	

20 25 30
Lys Leu Lys Ile Lys Ser Met Ser Gln Lys Lys Val Glu Ala Thr Ala
35 40 45
His His
50

(2) INFORMATION FOR SEQ ID NO:271:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 800 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..800
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481976

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:271:

```
atctatgcct acaccaacaa gcaacgggtca tgcctctcgc gtgcagattc aagaaccaag      60
aataatgtct cctcttcctc cttcttcttc tccaatcgcc ttcaaggaac aacaaggtag      120
accacctcca acaacacaa aaaccatagc aggaaaactc tttagaactc ttttcaaggg      180
tcttctcttc tcacaactaa ccttaatctc acttttggtg atcgttctca ccattcgcg      240
tctcatctca gcaagtacac accatttcca cctcaagaaa tggtagccctc ctttactagc      300
atctgttgct gtctcaggaa ttgcatcttt agcatggcaa tgcattctta tctacaatcc      360
atcaagagca gtcaaagcaa cgttctggct tagtccaata ctcacctgct cggtaggaat      420
cttgcttggt ttgattggct cagcggtaga tgcaggtata ggtgcagtgt ttgtcctttt      480
cgccattact cagtccctct atggttgctg gattactccg aggccttgagt acaccgataa      540
aatattatca cttgccacag catttccacc tgcaagaacc agagaagtag tctgcttatc      600
aatcatagtc agtgtcgttt actctggttt cttggtgact ggaattggag gagcaacttc      660
cactagaaca aatcttgata tcttgttcat atccgtaatc ataataagct tagcatggac      720
gatgcaagtt atcaagaatg ttcaacaagt tgcgatttca cgggcgagat atgtaaactt      780
tgcacatgga gaagatatgg
```

(2) INFORMATION FOR SEQ ID NO:272:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 266 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..266
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481977

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:272:

```
Ser Met Pro Thr Pro Thr Ser Asn Gly His Ala Ser Arg Val Gln Ile
1          5          10          15
Gln Glu Pro Arg Ile Met Ser Pro Leu Pro Pro Ser Ser Ser Pro Ile
20          25          30
Ala Phe Lys Glu Gln Gln Gly Arg Pro Pro Pro Thr Thr Gln Gln Thr
35          40          45
Ile Ala Gly Lys Leu Phe Arg Thr Leu Phe Lys Gly Leu Leu Phe Ser
50          55          60
Gln Leu Thr Leu Ile Ser Leu Leu Val Ile Val Leu Thr Ile Arg Gly
65          70          75          80
Leu Ile Ser Ala Ser Thr His His Phe His Leu Lys Lys Trp Tyr Pro
85          90          95
Pro Leu Leu Ala Ser Val Ala Val Ser Gly Ile Ala Ser Leu Ala Trp
100         105         110
Gln Cys Ile Phe Ile Tyr Asn Pro Ser Arg Ala Val Lys Ala Thr Phe
115         120         125
```

Trp Leu Ser Pro Ile Leu Thr Cys Ser Val Gly Ile Leu Leu Val Leu
130 135 140
Ile Gly Ser Ala Val Asp Ala Gly Ile Gly Ala Val Phe Val Leu Phe
145 150 155 160
Ala Ile Thr Gln Ser Leu Tyr Gly Cys Trp Ile Thr Pro Arg Leu Glu
165 170 175
Tyr Thr Asp Lys Ile Leu Ser Leu Ala Thr Ala Phe Pro Pro Ala Arg
180 185 190
Thr Arg Glu Val Val Cys Leu Ser Ile Ile Val Ser Val Val Tyr Ser
195 200 205
Gly Phe Leu Val Thr Gly Ile Gly Gly Ala Thr Ser Thr Arg Thr Asn
210 215 220
Leu Asp Ile Leu Phe Ile Ser Val Ile Ile Ile Ser Leu Ala Trp Thr
225 230 235 240
Met Gln Val Ile Lys Asn Val Gln Gln Val Ala Ile Ser Arg Ala Arg
245 250 255
Tyr Val Asn Phe Ala His Gly Glu Asp Met
260 265

(2) INFORMATION FOR SEQ ID NO:273:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 265 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..265
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481978

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:273:

Met Pro Thr Pro Thr Ser Asn Gly His Ala Ser Arg Val Gln Ile Gln
1 5 10 15
Glu Pro Arg Ile Met Ser Pro Leu Pro Pro Ser Ser Ser Pro Ile Ala
20 25 30
Phe Lys Glu Gln Gln Gly Arg Pro Pro Pro Thr Thr Gln Gln Thr Ile
35 40 45
Ala Gly Lys Leu Phe Arg Thr Leu Phe Lys Gly Leu Leu Phe Ser Gln
50 55 60
Leu Thr Leu Ile Ser Leu Leu Val Ile Val Leu Thr Ile Arg Gly Leu
65 70 75 80
Ile Ser Ala Ser Thr His His Phe His Leu Lys Lys Trp Tyr Pro Pro
85 90 95
Leu Leu Ala Ser Val Ala Val Ser Gly Ile Ala Ser Leu Ala Trp Gln
100 105 110
Cys Ile Phe Ile Tyr Asn Pro Ser Arg Ala Val Lys Ala Thr Phe Trp
115 120 125
Leu Ser Pro Ile Leu Thr Cys Ser Val Gly Ile Leu Leu Val Leu Ile
130 135 140
Gly Ser Ala Val Asp Ala Gly Ile Gly Ala Val Phe Val Leu Phe Ala
145 150 155 160
Ile Thr Gln Ser Leu Tyr Gly Cys Trp Ile Thr Pro Arg Leu Glu Tyr
165 170 175
Thr Asp Lys Ile Leu Ser Leu Ala Thr Ala Phe Pro Pro Ala Arg Thr
180 185 190
Arg Glu Val Val Cys Leu Ser Ile Ile Val Ser Val Val Tyr Ser Gly
195 200 205
Phe Leu Val Thr Gly Ile Gly Gly Ala Thr Ser Thr Arg Thr Asn Leu
210 215 220
Asp Ile Leu Phe Ile Ser Val Ile Ile Ile Ser Leu Ala Trp Thr Met

(2) INFORMATION FOR SEQ ID NO:274:

(A) LENGTH: 245 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..245

(D) OTHER INFORMATION: / Ceres Seq. ID 1481979

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:274:

(2) INFORMATION FOR SEQ ID NO:275:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 711 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..711

65

70

(2) INFORMATION FOR SEQ ID NO:278:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 750 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..750
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481983

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:278:

aacattacac	acagttcaag	aaagagatcg	atgtcgacct	tggaatctcc	attagaggct	60
ctggcgtttg	aatacgctag	cttcgggtgtt	ttcgccgtcg	tcaacaacgt	ctggacatgg	120
atcgccgtcg	tgactgccgc	cgtcagcttc	tggaggatca	gagtcacaac	catcggaagtc	180
ggagacggcc	atgcatgtgt	cttgcataaa	gaattaaccg	gttctaaatc	tgaaaacgaa	240
tccggtcgtc	tcgaaccaa	atcaataacc	ggcccggtca	aagaaacggt	tgcacgagtg	300
aaggaaacgg	ttacgaaaac	ggagccgtta	atatgcgatg	acggagtgac	aaagagggaag	360
ctgacgatgt	actacgaggt	agacgttgac	gttgacggtg	ggaggtgtgt	taacggagat	420
ttaacggcag	ttagctacgg	aggaggtttg	ggtaattgtg	gcgggggattg	gstggggagaa	480
atgggatgga	gtgggtgagga	tgagaaatgg	tgatgacagt	tggatccgtt	acgtggattt	540
aacggtgatt	aatggaaatg	tggttaaggt	atgggacgac	aacaaaacac	tagtaacggc	600
ggcatgtgtc	taaattagac	aagtttcata	tttcggaaa	tttttaaatac	tagagaaact	660
ttcttgcttt	aaagtgtttt	tttttttggt	tgattaagat	ctgtaatttg	taaataattt	720
tcacvrcaag	agaccaagaa	ggaacgcttg				

(2) INFORMATION FOR SEQ ID NO:279:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 170 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..170
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481984

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:279:

Asn	Ile	Thr	His	Ser	Ser	Arg	Lys	Arg	Ser	Met	Ser	Thr	Leu	Glu	Ser
1			5					10						15	
Pro	Leu	Glu	Ala	Leu	Ala	Phe	Glu	Tyr	Ala	Ser	Phe	Gly	Val	Phe	Ala
			20					25					30		
Val	Val	Asn	Asn	Val	Trp	Thr	Trp	Ile	Ala	Val	Val	Thr	Ala	Ala	Val
			35				40					45			
Ser	Phe	Trp	Arg	Ile	Arg	Val	Thr	Thr	Ile	Gly	Val	Gly	Asp	Gly	His
			50			55				60					
Ala	Cys	Val	Leu	Ile	Glu	Glu	Leu	Thr	Gly	Ser	Lys	Ser	Glu	Asn	Glu
65					70				75					80	
Ser	Gly	Arg	Leu	Glu	Pro	Lys	Ser	Ile	Thr	Gly	Pro	Val	Lys	Glu	Thr
			85					90						95	
Val	Ala	Arg	Val	Lys	Glu	Thr	Val	Thr	Lys	Thr	Glu	Pro	Leu	Ile	Cys
			100					105					110		
Asp	Asp	Gly	Val	Thr	Lys	Arg	Lys	Leu	Thr	Met	Tyr	Tyr	Glu	Val	Asp
			115				120					125			
Val	Asp	Val	Asp	Gly	Gly	Arg	Cys	Val	Asn	Gly	Asp	Leu	Thr	Ala	Val
			130			135					140				
Ser	Tyr	Gly	Gly	Gly	Leu	Gly	Asn	Cys	Gly	Gly	Asp	Trp	Xaa	Gly	Glu
145					150				155						160
Met	Gly	Trp	Ser	Gly	Glu	Asp	Glu	Lys	Trp						

165 170

(2) INFORMATION FOR SEQ ID NO:280:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 160 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..160

(D) OTHER INFORMATION: / Ceres Seq. ID 1481985

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:280:

Met	Ser	Thr	Leu	Glu	Ser	Pro	Leu	Glu	Ala	Leu	Ala	Phe	Glu	Tyr	Ala
1				5					10					15	
Ser	Phe	Gly	Val	Phe	Ala	Val	Val	Asn	Val	Trp	Thr	Trp	Ile	Ala	
			20					25				30			
Val	Val	Thr	Ala	Ala	Val	Ser	Phe	Trp	Arg	Ile	Arg	Val	Thr	Thr	Ile
			35					40				45			
Gly	Val	Gly	Asp	Gly	His	Ala	Cys	Val	Leu	Ile	Glu	Glu	Leu	Thr	Gly
			50				55				60				
Ser	Lys	Ser	Glu	Asn	Glu	Ser	Gly	Arg	Leu	Glu	Pro	Lys	Ser	Ile	Thr
65					70				75					80	
Gly	Pro	Val	Lys	Glu	Thr	Val	Ala	Arg	Val	Lys	Glu	Thr	Val	Thr	Lys
			85					90						95	
Thr	Glu	Pro	Leu	Ile	Cys	Asp	Asp	Gly	Val	Thr	Lys	Arg	Lys	Leu	Thr
			100					105						110	
Met	Tyr	Tyr	Glu	Val	Asp	Val	Asp	Val	Asp	Gly	Gly	Arg	Cys	Val	Asn
			115					120					125		
Gly	Asp	Leu	Thr	Ala	Val	Ser	Tyr	Gly	Gly	Gly	Leu	Gly	Asn	Cys	Gly
			130				135				140				
Gly	Asp	Trp	Xaa	Gly	Glu	Met	Gly	Trp	Ser	Gly	Glu	Asp	Glu	Lys	Trp
145					150					155					160

(2) INFORMATION FOR SEQ ID NO:281:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 598 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..598

(D) OTHER INFORMATION: / Ceres Seq. ID 1481986

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:281:

gaaaaggagc	ccttcttcaa	aattgggtca	tgtactcatg	cttcttcttc	ttcttagctt	60
cctattgcac	cataccgaat	ctactttgcc	tcctgatcat	gaacaactct	caataaatgg	120
gaggagaatt	atggcgat	acaagcacga	tggtgccata	gcagcaccac	catcaagaag	180
tggtgagagt	gggtgtcacg	ggaagaggat	gatgccctac	cataagccaa	atgctcctat	240
acaaacacca	ccatcaagaa	gtagacgacg	tgagggtggt	cacaacggga	gtagacagat	300
gggtatatat	aggccaaatg	gagacatata	tacaggacca	tcaaatagtg	gacatggtgg	360
tggtcacatt	catcaaaatt	catctcctta	gttttggggc	aatttacaaa	attggaaact	420
tatctaaaaa	ttcgccaaaa	agattataga	tttgaatgta	atttgtgttt	catgtgatcc	480
caagtatgga	gtggatatgg	tggtgggtcac	attcatcaac	atttcgatct	ccttagtttt	540
ataygatatg	aatgtaattg	tattttatgt	tattccaagt	aaggatatat	aaagtcgc	

(2) INFORMATION FOR SEQ ID NO:282:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 129 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
(A) NAME/KEY: peptide
(B) LOCATION: 1..129
(D) OTHER INFORMATION: / Ceres Seq. ID 1481987
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:282:

Lys	Arg	Ser	Pro	Ser	Ser	Lys	Leu	Gly	His	Val	Leu	Met	Leu	Leu	Leu
1				5				10					15		
Leu	Leu	Ser	Phe	Leu	Leu	His	His	Thr	Glu	Ser	Thr	Leu	Pro	Pro	Asp
			20					25				30			
His	Glu	Gln	Leu	Ser	Ile	Asn	Gly	Arg	Arg	Ile	Met	Ala	Tyr	Tyr	Lys
		35				40					45				
His	Asp	Gly	Ala	Ile	Ala	Ala	Pro	Pro	Ser	Arg	Ser	Gly	Arg	Gly	Gly
	50					55				60					
Gly	His	Gly	Lys	Arg	Met	Met	Pro	Tyr	His	Lys	Pro	Asn	Ala	Pro	Ile
65					70				75					80	
Gln	Thr	Pro	Pro	Ser	Arg	Ser	Arg	Arg	Arg	Glu	Gly	Gly	His	Asn	Gly
				85				90						95	
Ser	Arg	Gln	Met	Gly	Ile	Tyr	Arg	Pro	Asn	Gly	Asp	Ile	Tyr	Thr	Gly
		100						105				110			
Pro	Ser	Asn	Ser	Gly	His	Gly	Gly	Gly	His	Ile	His	Gln	Asn	Ser	Ser
		115				120						125			
Pro															

(2) INFORMATION FOR SEQ ID NO:283:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 117 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide
(B) LOCATION: 1..117

(D) OTHER INFORMATION: / Ceres Seq. ID 1481988

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:283:

Met	Leu	Leu	Leu	Leu	Ser	Phe	Leu	Leu	His	His	Thr	Glu	Ser	Thr	
1				5				10				15			
Leu	Pro	Pro	Asp	His	Glu	Gln	Leu	Ser	Ile	Asn	Gly	Arg	Arg	Ile	Met
			20					25				30			
Ala	Tyr	Tyr	Lys	His	Asp	Gly	Ala	Ile	Ala	Ala	Pro	Pro	Ser	Arg	Ser
		35				40					45				
Gly	Arg	Gly	Gly	Gly	His	Gly	Lys	Arg	Met	Met	Pro	Tyr	His	Lys	Pro
	50					55				60					
Asn	Ala	Pro	Ile	Gln	Thr	Pro	Pro	Ser	Arg	Ser	Arg	Arg	Arg	Glu	Gly
65				70					75					80	
Gly	His	Asn	Gly	Ser	Arg	Gln	Met	Gly	Ile	Tyr	Arg	Pro	Asn	Gly	Asp
			85					90					95		
Ile	Tyr	Thr	Gly	Pro	Ser	Asn	Ser	Gly	His	Gly	Gly	Gly	His	Ile	His
		100						105					110		
Gln	Asn	Ser	Ser	Pro											
		115													

(2) INFORMATION FOR SEQ ID NO:284:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 86 amino acids

- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..86
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1481989
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:284:

```
Met Ala Tyr Tyr Lys His Asp Gly Ala Ile Ala Ala Pro Pro Ser Arg
1           5           10           15
Ser Gly Arg Gly Gly Gly His Gly Lys Arg Met Met Pro Tyr His Lys
          20           25           30
Pro Asn Ala Pro Ile Gln Thr Pro Pro Ser Arg Ser Arg Arg Arg Glu
          35           40           45
Gly Gly His Asn Gly Ser Arg Gln Met Gly Ile Tyr Arg Pro Asn Gly
          50           55           60
Asp Ile Tyr Thr Gly Pro Ser Asn Ser Gly His Gly Gly Gly His Ile
65           70           75           80
His Gln Asn Ser Ser Pro
          85
```

(2) INFORMATION FOR SEQ ID NO:285:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 688 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- (ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..688
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481990

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:285:

```
gtggtattac cgaacttaaa cctcgctgctc gagcttcgaa actcttttttc tcagttcacc      60
tggaaaacga tgcgctcgta caagcagacg ccatgagagc ttgcaggagt ctctggagaa      120
atttgagat ttagatgaac tccaatctat ctgctgatga taatgataag gatttggaat      180
aactataagg gtaaatacag attcttcctc tcaaattgcc gctctttctc gtcaatcaaa      240
cgaccccaaa tcccagaaag cgaagagact agcctctcga tcacacaacg aagattcgac      300
ccagatttag ctccatcaaa gactagagtt tacgtctctc tcttccatac tctctttcgg      360
ctctatttaa gctgtgagag actctacgga gcagcaagga cgctctctgc gatgtgcact      420
ttcgggggtg ttccggattc gcgtctatgg aacagtctga ttcatacaatt caatgtcaat      480
ggtttggtac acgatcaggt atcgctgatt tacagcaaga tgatagcttg tggagtttct      540
cccgatgttt ttgctctcaa tgtattgatt cattcttttt gcaaagtggg tcgtttgagt      600
tttgcaatta gtttacttag aaatagagta atcagcatcg atactgttac ttataacact      660
gtgatttcgg gtttatgtga acatggct
```

(2) INFORMATION FOR SEQ ID NO:286:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 177 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..177
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481991

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:286:

```
Met Ile Met Ile Arg Ile Trp Asn Asn Tyr Lys Gly Lys Tyr Arg Phe
1           5           10           15
```

Phe Leu Ser Asn Cys Arg Ser Phe Ser Ser Ile Lys Arg Pro Gln Ile
20 25 30
Pro Glu Ser Glu Glu Thr Ser Leu Ser Ile Thr Gln Arg Arg Phe Asp
35 40 45
Pro Asp Leu Ala Pro Ile Lys Thr Arg Val Tyr Val Ser Leu Phe His
50 55 60
Thr Leu Phe Arg Leu Tyr Leu Ser Cys Glu Arg Leu Tyr Gly Ala Ala
65 70 75 80
Arg Thr Leu Ser Ala Met Cys Thr Phe Gly Val Val Pro Asp Ser Arg
85 90 95
Leu Trp Asn Ser Leu Ile His Gln Phe Asn Val Asn Gly Leu Val His
100 105 110
Asp Gln Val Ser Leu Ile Tyr Ser Lys Met Ile Ala Cys Gly Val Ser
115 120 125
Pro Asp Val Phe Ala Leu Asn Val Leu Ile His Ser Phe Cys Lys Val
130 135 140
Gly Arg Leu Ser Phe Ala Ile Ser Leu Leu Arg Asn Arg Val Ile Ser
145 150 155 160
Ile Asp Thr Val Thr Tyr Asn Thr Val Ile Ser Gly Leu Cys Glu His
165 170 175
Gly

(2) INFORMATION FOR SEQ ID NO:287:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 175 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..175
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481992

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:287:

Met Ile Arg Ile Trp Asn Asn Tyr Lys Gly Lys Tyr Arg Phe Phe Leu
1 5 10 15
Ser Asn Cys Arg Ser Phe Ser Ser Ile Lys Arg Pro Gln Ile Pro Glu
20 25 30
Ser Glu Glu Thr Ser Leu Ser Ile Thr Gln Arg Arg Phe Asp Pro Asp
35 40 45
Leu Ala Pro Ile Lys Thr Arg Val Tyr Val Ser Leu Phe His Thr Leu
50 55 60
Phe Arg Leu Tyr Leu Ser Cys Glu Arg Leu Tyr Gly Ala Ala Arg Thr
65 70 75 80
Leu Ser Ala Met Cys Thr Phe Gly Val Val Pro Asp Ser Arg Leu Trp
85 90 95
Asn Ser Leu Ile His Gln Phe Asn Val Asn Gly Leu Val His Asp Gln
100 105 110
Val Ser Leu Ile Tyr Ser Lys Met Ile Ala Cys Gly Val Ser Pro Asp
115 120 125
Val Phe Ala Leu Asn Val Leu Ile His Ser Phe Cys Lys Val Gly Arg
130 135 140
Leu Ser Phe Ala Ile Ser Leu Leu Arg Asn Arg Val Ile Ser Ile Asp
145 150 155 160
Thr Val Thr Tyr Asn Thr Val Ile Ser Gly Leu Cys Glu His Gly
165 170 175

(2) INFORMATION FOR SEQ ID NO:288:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 92 amino acids

(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
(A) NAME/KEY: peptide
(B) LOCATION: 1..92
(D) OTHER INFORMATION: / Ceres Seq. ID 1481993
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:288:
Met Cys Thr Phe Gly Val Val Pro Asp Ser Arg Leu Trp Asn Ser Leu
1 5 10 15
Ile His Gln Phe Asn Val Asn Gly Leu Val His Asp Gln Val Ser Leu
20 25 30
Ile Tyr Ser Lys Met Ile Ala Cys Gly Val Ser Pro Asp Val Phe Ala
35 40 45
Leu Asn Val Leu Ile His Ser Phe Cys Lys Val Gly Arg Leu Ser Phe
50 55 60
Ala Ile Ser Leu Leu Arg Asn Arg Val Ile Ser Ile Asp Thr Val Thr
65 70 75 80
Tyr Asn Thr Val Ile Ser Gly Leu Cys Glu His Gly
85 90

(2) INFORMATION FOR SEQ ID NO:289:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 499 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
(B) LOCATION: 1..499
(D) OTHER INFORMATION: / Ceres Seq. ID 1481994

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:289:

attgtaactt gtaaccagtg tcggctaatt tcgacttttg tagatctttt tctgctcttt 60
ctctctctct ctgctctctc tctctctctc tctctcttgt attatttcta tctccccgcg 120
cgctcgaaaga gaaacgtcga tcggagaacc tttgaaatgt cgactggatt agatatgtct 180
ctcgacgaca tgatcgccaa gaaccgtaag tctcgtgggtg gagccggccc cgctcgtgga 240
accggatccg gatccggacc gggtcggact cgccgcaaca accctaatacg gaaatcaacc 300
cgatctgctc cataccaatc agccaaggcg ccggagtcca cctgggggtca cgacatgttc 360
tccgatagat ctgaagatca ccgatcggga cgttcctccg ccggaatcga aactggaacc 420
aagctctaca tttccaattt ggayttacgg tgtcatgaac gaagacatca aggaactggt 480
tgctgaagtt ggagaactt

(2) INFORMATION FOR SEQ ID NO:290:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 161 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
(B) LOCATION: 1..161
(D) OTHER INFORMATION: / Ceres Seq. ID 1481995

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:290:

Ile Val Thr Cys Asn Gln Cys Arg Leu Ile Ser Thr Phe Val Asp Leu
1 5 10 15
Phe Leu Leu Phe Leu Ser Leu Ser Ala Leu Ser Leu Ser Leu Ser Leu
20 25 30
Leu Tyr Tyr Phe Tyr Leu Pro Arg Arg Arg Lys Arg Asn Val Asp Arg

35	40	45
Arg Thr Phe Glu Met Ser	Thr Gly Leu Asp Met	Ser Leu Asp Asp Met
50	55	60
Ile Ala Lys Asn Arg Lys	Ser Arg Gly Gly Ala	Gly Pro Ala Arg Gly
65	70	75
Thr Gly Ser Gly Ser Gly	Pro Gly Pro Thr Arg	Arg Asn Asn Pro Asn
85	90	95
Arg Lys Ser Thr Arg Ser	Ala Pro Tyr Gln Ser	Ala Lys Ala Pro Glu
100	105	110
Ser Thr Trp Gly His Asp	Met Phe Ser Asp Arg	Ser Glu Asp His Arg
115	120	125
Ser Gly Arg Ser Ser Ala	Gly Ile Glu Thr Gly	Thr Lys Leu Tyr Ile
130	135	140
Ser Asn Leu Xaa Leu Arg	Cys His Glu Arg Arg	His Gln Gly Thr Val
145	150	155
Cys		160

(2) INFORMATION FOR SEQ ID NO:291:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 109 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..109
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481996

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:291:

Met Ser Thr Gly Leu Asp Met Ser Leu Asp Asp Met Ile Ala Lys Asn	
1	15
Arg Lys Ser Arg Gly Gly Ala Gly Pro Ala Arg Gly Thr Gly Ser Gly	
20	30
Ser Gly Pro Gly Pro Thr Arg Arg Asn Asn Pro Asn Arg Lys Ser Thr	
35	45
Arg Ser Ala Pro Tyr Gln Ser Ala Lys Ala Pro Glu Ser Thr Trp Gly	
50	60
His Asp Met Phe Ser Asp Arg Ser Glu Asp His Arg Ser Gly Arg Ser	
65	80
Ser Ala Gly Ile Glu Thr Gly Thr Lys Leu Tyr Ile Ser Asn Leu Xaa	
85	95
Leu Arg Cys His Glu Arg Arg His Gln Gly Thr Val Cys	
100	105

(2) INFORMATION FOR SEQ ID NO:292:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 103 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..103
- (D) OTHER INFORMATION: / Ceres Seq. ID 1481997

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:292:

Met Ser Leu Asp Asp Met Ile Ala Lys Asn Arg Lys Ser Arg Gly Gly	
1	15
Ala Gly Pro Ala Arg Gly Thr Gly Ser Gly Ser Gly Pro Gly Pro Thr	
20	30

Arg Arg Asn Asn Pro Asn Arg Lys Ser Thr Arg Ser Ala Pro Tyr Gln
35 40 45
Ser Ala Lys Ala Pro Glu Ser Thr Trp Gly His Asp Met Phe Ser Asp
50 55 60
Arg Ser Glu Asp His Arg Ser Gly Arg Ser Ser Ala Gly Ile Glu Thr
65 70 75 80
Gly Thr Lys Leu Tyr Ile Ser Asn Leu Xaa Leu Arg Cys His Glu Arg
85 90 95
Arg His Gln Gly Thr Val Cys
100

(2) INFORMATION FOR SEQ ID NO:293:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 851 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..851
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482009

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:293:

agagagaatc	gcattaacaa	aaaaacaaac	gaatcttttg	agtttaaaac	cctttttcac	60
ttaccggaga	aatggagaga	tcgacgccgg	aacatgtctc	ctccgcacac	aagcgcataa	120
gcgtgagctt	cctcgtgtct	ctcatggtag	tttgtgctag	acacgcaagc	agagtttcca	180
agaagcttaa	acccaagaag	actcggaagc	aaactcatct	tgaagactat	ctcgaaagcc	240
ctaagtctaa	cggaaacggt	agcgaagacg	gtagaggagg	aggaagggtt	ggatggagtc	300
cggcaaggac	tttttctcct	atgagggtgc	gtcctaagga	gctctacacg	accttgagca	360
acaaggcgat	gactatgggt	ggccggaaaa	acaaagctta	cgacggtggt	ccgacgaaga	420
agacggcggt	ggagatgggt	atggaggagg	atgaggaaga	gtacggcggt	tggcagaggg	480
agattttgat	gggaggaaaa	tgtgagccgt	tggattactc	aggcgtgatc	tactacgatt	540
gtagtggaca	tcagctaaaa	caagtgcctc	caaggtctcc	acgtgccagt	ttggttccgg	600
agcgcgccag	tcgttcttat	gtcgggtcat	tgtaaacc	gacgggaaag	gaaatttaa	660
ttttagtttg	agaatttgaa	attttagtag	gagtatttga	ttgttggttg	aggtgtcatc	720
acgtaagtgg	taaattctct	aggagctttg	ttggctccct	gtcattagta	gatgcatgac	780
atgtttttat	gcataattgt	gtgtagttta	tgtatttaag	acgtttggca	attttaaaac	840
tttagtagtt	t					

(2) INFORMATION FOR SEQ ID NO:294:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 195 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..195
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482010

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:294:

Met	Glu	Arg	Ser	Thr	Pro	Glu	His	Val	Ser	Ser	Ala	His	Lys	Arg	Ile
1				5						10				15	
Ser	Val	Ser	Phe	Leu	Val	Ser	Leu	Met	Val	Leu	Cys	Ala	Arg	His	Ala
			20					25					30		
Ser	Arg	Val	Ser	Lys	Lys	Leu	Lys	Pro	Lys	Lys	Thr	Arg	Lys	Gln	Thr
			35				40				45				
His	Leu	Glu	Asp	Tyr	Leu	Glu	Ser	Pro	Lys	Ser	Asn	Gly	Asn	Gly	Ser
	50					55				60					
Glu	Asp	Gly	Arg	Gly	Gly	Gly	Arg	Phe	Gly	Trp	Ser	Pro	Ala	Arg	Thr
65				70					75					80	

Phe Ser Pro Met Arg Val Arg Pro Lys Glu Leu Tyr Thr Thr Leu Ser
85 90 95
Asn Lys Ala Met Thr Met Val Gly Arg Lys Asn Lys Ala Tyr Asp Gly
100 105 110
Gly Pro Thr Lys Lys Thr Ala Val Glu Met Val Met Glu Glu Asp Glu
115 120 125
Glu Glu Tyr Gly Val Trp Gln Arg Glu Ile Leu Met Gly Gly Lys Cys
130 135 140
Glu Pro Leu Asp Tyr Ser Gly Val Ile Tyr Tyr Asp Cys Ser Gly His
145 150 155 160
Gln Leu Lys Gln Val Pro Pro Arg Ser Pro Arg Ala Ser Leu Val Pro
165 170 175
Glu Arg Pro Thr Arg Ser Tyr Val Gly Ser Leu Leu Asn Pro Thr Gly
180 185 190
Lys Glu Ile
195

(2) INFORMATION FOR SEQ ID NO:295:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 171 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..171
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482011

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:295:

Met Val Leu Cys Ala Arg His Ala Ser Arg Val Ser Lys Lys Leu Lys
1 5 10 15
Pro Lys Lys Thr Arg Lys Gln Thr His Leu Glu Asp Tyr Leu Glu Ser
20 25 30
Pro Lys Ser Asn Gly Asn Gly Ser Glu Asp Gly Arg Gly Gly Arg
35 40 45
Phe Gly Trp Ser Pro Ala Arg Thr Phe Ser Pro Met Arg Val Arg Pro
50 55 60
Lys Glu Leu Tyr Thr Thr Leu Ser Asn Lys Ala Met Thr Met Val Gly
65 70 75 80
Arg Lys Asn Lys Ala Tyr Asp Gly Gly Pro Thr Lys Lys Thr Ala Val
85 90 95
Glu Met Val Met Glu Glu Asp Glu Glu Tyr Gly Val Trp Gln Arg
100 105 110
Glu Ile Leu Met Gly Gly Lys Cys Glu Pro Leu Asp Tyr Ser Gly Val
115 120 125
Ile Tyr Tyr Asp Cys Ser Gly His Gln Leu Lys Gln Val Pro Pro Arg
130 135 140
Ser Pro Arg Ala Ser Leu Val Pro Glu Arg Pro Thr Arg Ser Tyr Val
145 150 155 160
Gly Ser Leu Leu Asn Pro Thr Gly Lys Glu Ile
165 170

(2) INFORMATION FOR SEQ ID NO:296:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 112 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide

(B) LOCATION: 1..112

(D) OTHER INFORMATION: / Ceres Seq. ID 1482012

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:296:

```
Met Arg Val Arg Pro Lys Glu Leu Tyr Thr Thr Leu Ser Asn Lys Ala
1          5          10          15
Met Thr Met Val Gly Arg Lys Asn Lys Ala Tyr Asp Gly Gly Pro Thr
          20          25          30
Lys Lys Thr Ala Val Glu Met Val Met Glu Glu Asp Glu Glu Glu Tyr
          35          40          45
Gly Val Trp Gln Arg Glu Ile Leu Met Gly Gly Lys Cys Glu Pro Leu
          50          55          60
Asp Tyr Ser Gly Val Ile Tyr Tyr Asp Cys Ser Gly His Gln Leu Lys
65          70          75          80
Gln Val Pro Pro Arg Ser Pro Arg Ala Ser Leu Val Pro Glu Arg Pro
          85          90          95
Thr Arg Ser Tyr Val Gly Ser Leu Leu Asn Pro Thr Gly Lys Glu Ile
          100          105          110
```

(2) INFORMATION FOR SEQ ID NO:297:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 576 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..576
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482013

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:297:

```
agatttgcac tgcgagggga taaggatcaa aaatggagga agcaaaggga cctgtgaagc      60
acgtattgct tgctagtttc aaagatgggg ttagtcctga gaaaatcgaa gagctcatca      120
aaggttacgc caatctcgtc aatctcatcg aacctatgaa agctttccac tggggaaaag      180
atgtgagcat tgagaatctg catcaagggt acacacacat ctttgaatcc acatttgaga      240
gtaaagaagc tgttgagagc tacattgctc atcctgctca cgttaaattc gccaccatct      300
tccttggcag cttggataaa gttttgggta ttgactacaa gcctacctct gtctctctct      360
aattatcttg tagcagcatt ttcattcatt atctttttct cgggtatgca tcttgtatgt      420
tgaataaagt atattccttt tgagttttcc tgcattgttc tcatgtttct ctgtgaattt      480
ctctcttttt tgtttgtttg tttgtttcct tctgttgtat tatacttgat ctgtaaaaag      540
atcatgagtt tattaagagt gtttgatttc agactc
```

(2) INFORMATION FOR SEQ ID NO:298:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 109 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..109
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482014

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:298:

```
Met Glu Glu Ala Lys Gly Pro Val Lys His Val Leu Leu Ala Ser Phe
1          5          10          15
Lys Asp Gly Val Ser Pro Glu Lys Ile Glu Glu Leu Ile Lys Gly Tyr
          20          25          30
Ala Asn Leu Val Asn Leu Ile Glu Pro Met Lys Ala Phe His Trp Gly
          35          40          45
```

Lys Asp Val Ser Ile Glu Asn Leu His Gln Gly Tyr Thr His Ile Phe
50 55 60
Glu Ser Thr Phe Glu Ser Lys Glu Ala Val Ala Glu Tyr Ile Ala His
65 70 75 80
Pro Ala His Val Lys Phe Ala Thr Ile Phe Leu Gly Ser Leu Asp Lys
85 90 95
Val Leu Val Ile Asp Tyr Lys Pro Thr Ser Val Ser Leu
100 105

(2) INFORMATION FOR SEQ ID NO:299:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 68 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..68
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482015

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:299:

Met Lys Ala Phe His Trp Gly Lys Asp Val Ser Ile Glu Asn Leu His
1 5 10 15
Gln Gly Tyr Thr His Ile Phe Glu Ser Thr Phe Glu Ser Lys Glu Ala
20 25 30
Val Ala Glu Tyr Ile Ala His Pro Ala His Val Lys Phe Ala Thr Ile
35 40 45
Phe Leu Gly Ser Leu Asp Lys Val Leu Val Ile Asp Tyr Lys Pro Thr
50 55 60
Ser Val Ser Leu
65

(2) INFORMATION FOR SEQ ID NO:300:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 664 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..664
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482016

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:300:

gattaat ttt tgagagagct gtctctcttg acagagattt tggaaggtaa gagagacgat 60
gacgtatcac gtttttagac gagactatgg cgatggagag tgaattttgc ctttttgga 120
ttcccacgac tctctgtatc tttcttttagg cgagactatg gcgataaaga ttgaattttg 180
cttcgacaat tgagggtgaa attaacggca aattcaaaat tgcggtttct gacaagtcgt 240
tccgctggat tcgtcagctt tctcaccatc ggcggaagcga ggttcaccta atcgagatt 300
tgcgactccc agttggagag taatcggtga ggagaggcaa cgagtgcacac gagcatatca 360
cttctctcgc cattcttcgt acccatcgca agctaggtct cgtcactaag ctcagtgtg 420
ccgctcaggc tgctatggaa cagcttaaaag gtatgataaa cgacatggat cgtgtccaac 480
tggaatgagg actcttgtgt gttacaccta tcgtcaatgc ccaactgata tgttgtgtct 540
tataaccata aatttacttt gatccaaaca cttttgagaa gctgtcttca agtggtcaaa 600
aggtagcaac tcttttttct tgtgtaattg taatcatctg tgttatgaag tattgccatt 660
ttcg

(2) INFORMATION FOR SEQ ID NO:301:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 958 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single

- (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: DNA (genomic)
(ix) FEATURE:
(A) NAME/KEY: -
(B) LOCATION: 1..958
(D) OTHER INFORMATION: / Ceres Seq. ID 1482021

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:301:

aaagccctaa	aaatcagaga	ttccattttc	tcttatctct	ctctctctct	ctttctcttt	60
ttccgattct	gattctatct	tttcttcacc	aaccacacaa	aacaattcta	cgtttgatct	120
cttcttcttt	ctccgtccaa	attaatctct	acgtttaatt	tctcttggtc	aatcatggga	180
cacgaaacaa	tgacgccggc	aacaacaacg	ctcgtgttca	cgtacggaac	tctaaagaga	240
ggattctcaa	atcatgtcct	gatgcaagat	ctgatccgat	ctgggtgacgc	ttctttcaaa	300
ggtgtttacc	aaactctaga	caaatatcct	ctcgtctgtg	gaccttaccg	agtccctttc	360
ctcctcaaca	aacctggatc	gggctatcac	gtcaccggcg	agctttacgc	ggtttctcct	420
cgcggtctct	ctcgtctcga	tgagcttgaa	ggaatcagtc	gcggtcatta	catccggcaa	480
ccgatacggt	ctcgcggcgg	cggaggaaga	agaagaagaa	ggagatctgg	aaacagaggc	540
gccgtcgtcg	tgcggtgtgg	aggcgtatta	cgctcataag	agttatgagg	aagagctgtg	600
gaagaggaat	agaggaagat	cattcggcgc	gtacacggaa	aacgaagcgc	gtggatatgt	660
gaaacgcaat	gataggcctc	agcatcttag	cttcttggtg	catatccgta	ttttcgtatc	720
ttctccatgt	gattgatttt	tatttctttc	gtgggtctct	ccgctcgtcg	cttttctatg	780
tttgtttgtt	tttttctcgg	gacaaaagaa	acaaaaaaaa	aacacaaaca	caaactagtt	840
ttacaacttg	taagggtccc	accagtccgt	ccgtccgctg	tctccgtatc	gatttgatta	900
gagagattgt	tgggtgtaaa	acttatgatt	cccatcttaa	ataagtttta	ggttgttt	

(2) INFORMATION FOR SEQ ID NO:302:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 134 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
(B) LOCATION: 1..134
(D) OTHER INFORMATION: / Ceres Seq. ID 1482022

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:302:

Met	Gly	His	Glu	Thr	Met	Thr	Pro	Ala	Thr	Thr	Thr	Leu	Val	Phe	Thr
1			5					10						15	
Tyr	Gly	Thr	Leu	Lys	Arg	Gly	Phe	Ser	Asn	His	Val	Leu	Met	Gln	Asp
			20					25						30	
Leu	Ile	Arg	Ser	Gly	Asp	Ala	Ser	Phe	Lys	Gly	Val	Tyr	Gln	Thr	Leu
			35					40						45	
Asp	Lys	Tyr	Pro	Leu	Val	Cys	Gly	Pro	Tyr	Arg	Val	Pro	Phe	Leu	Leu
			50					55						60	
Asn	Lys	Pro	Gly	Ser	Gly	Tyr	His	Val	Thr	Gly	Glu	Leu	Tyr	Ala	Val
								70						80	
Ser	Pro	Arg	Gly	Leu	Ser	Arg	Leu	Asp	Glu	Leu	Glu	Gly	Ile	Ser	Arg
								85						95	
Gly	His	Tyr	Ile	Arg	Gln	Pro	Ile	Arg	Ser	Arg	Gly	Gly	Gly	Gly	Arg
								105						110	
Arg	Arg	Arg	Arg	Ser	Gly	Asn	Arg	Gly	Ala	Val	Val	Val	Arg	Gly	
								120						125	
Gly	Gly	Val	Leu	Arg	Ser										

(2) INFORMATION FOR SEQ ID NO:303:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 129 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..129
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1482023

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:303:

```
Met Thr Pro Ala Thr Thr Leu Val Phe Thr Tyr Gly Thr Leu Lys
1      5      10      15
Arg Gly Phe Ser Asn His Val Leu Met Gln Asp Leu Ile Arg Ser Gly
      20      25      30
Asp Ala Ser Phe Lys Gly Val Tyr Gln Thr Leu Asp Lys Tyr Pro Leu
      35      40      45
Val Cys Gly Pro Tyr Arg Val Pro Phe Leu Leu Asn Lys Pro Gly Ser
      50      55      60
Gly Tyr His Val Thr Gly Glu Leu Tyr Ala Val Ser Pro Arg Gly Leu
65      70      75      80
Ser Arg Leu Asp Glu Leu Glu Gly Ile Ser Arg Gly His Tyr Ile Arg
      85      90      95
Gln Pro Ile Arg Ser Arg Gly Gly Gly Gly Arg Arg Arg Arg Arg Arg
      100      105      110
Ser Gly Asn Arg Gly Ala Val Val Val Arg Gly Gly Gly Val Leu Arg
      115      120      125
Ser
```

(2) INFORMATION FOR SEQ ID NO:304:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 105 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..105
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1482024

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:304:

```
Met Gln Asp Leu Ile Arg Ser Gly Asp Ala Ser Phe Lys Gly Val Tyr
1      5      10      15
Gln Thr Leu Asp Lys Tyr Pro Leu Val Cys Gly Pro Tyr Arg Val Pro
      20      25      30
Phe Leu Leu Asn Lys Pro Gly Ser Gly Tyr His Val Thr Gly Glu Leu
      35      40      45
Tyr Ala Val Ser Pro Arg Gly Leu Ser Arg Leu Asp Glu Leu Glu Gly
      50      55      60
Ile Ser Arg Gly His Tyr Ile Arg Gln Pro Ile Arg Ser Arg Gly Gly
65      70      75      80
Gly Gly Arg Arg Arg Arg Arg Arg Ser Gly Asn Arg Gly Ala Val Val
      85      90      95
Val Arg Gly Gly Gly Val Leu Arg Ser
      100      105
```

(2) INFORMATION FOR SEQ ID NO:305:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 535 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- (ix) FEATURE:
 - (A) NAME/KEY: -

(B) LOCATION: 1..535

(D) OTHER INFORMATION: / Ceres Seq. ID 1482029

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:305:

atcataactct	ctcaacttca	tctctctctc	tctctcaatc	tcttaagatc	ccacaagtca	60
cttttcttct	tcttaatcac	ctttaatggc	gaatttgatc	cttaagcaat	ctctaatacat	120
actcctaata	atatattcaa	caccaatctt	gagttctcaa	gctcgaatcc	tccgtacata	180
tcgccccaca	accatgggcg	atatggatag	tcaggttctc	ctacgtgaac	tcgggattga	240
tctctctaag	ttcaaaggtc	aagacgagag	acggttttta	gtggattccg	aaagggtttc	300
tccggggggg	cctgatccac	aacaccattg	actgatcttt	accgatatat	atatacttta	360
ccgaagatcg	aagcacacat	ataactgtga	ctgatccatg	caagtcaatt	taaatatcgt	420
catttacatg	cttttcttdt	ctttttcata	aatcttccct	acacttttgt	tgtatcaaga	480
ttttgggtatt	ctttwgtacc	ttccttatct	ttaaacaatca	aggttttact	ochtt	

(2) INFORMATION FOR SEQ ID NO:306:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 81 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..81

(D) OTHER INFORMATION: / Ceres Seq. ID 1482030

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:306:

Met	Ala	Asn	Leu	Ile	Leu	Lys	Gln	Ser	Leu	Ile	Ile	Leu	Leu	Ile	Ile
1			5				10					15			
Tyr	Ser	Thr	Pro	Ile	Leu	Ser	Ser	Gln	Ala	Arg	Ile	Leu	Arg	Thr	Tyr
			20				25					30			
Arg	Pro	Thr	Thr	Met	Gly	Asp	Met	Asp	Ser	Gln	Val	Leu	Leu	Arg	Glu
			35				40					45			
Leu	Gly	Ile	Asp	Leu	Ser	Lys	Phe	Lys	Gly	Gln	Asp	Glu	Arg	Arg	Phe
			50				55					60			
Leu	Val	Asp	Ser	Glu	Arg	Val	Ser	Pro	Gly	Gly	Pro	Asp	Pro	Gln	His
							70				75				80
His															

(2) INFORMATION FOR SEQ ID NO:307:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 45 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..45

(D) OTHER INFORMATION: / Ceres Seq. ID 1482031

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:307:

Met	Gly	Asp	Met	Asp	Ser	Gln	Val	Leu	Leu	Arg	Glu	Leu	Gly	Ile	Asp
1			5				10					15			
Leu	Ser	Lys	Phe	Lys	Gly	Gln	Asp	Glu	Arg	Arg	Phe	Leu	Val	Asp	Ser
			20				25					30			
Glu	Arg	Val	Ser	Pro	Gly	Gly	Pro	Asp	Pro	Gln	His	His			
			35				40					45			

(2) INFORMATION FOR SEQ ID NO:308:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 42 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

- (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
(A) NAME/KEY: peptide
(B) LOCATION: 1..42
(D) OTHER INFORMATION: / Ceres Seq. ID 1482032

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:308:

```
Met Asp Ser Gln Val Leu Leu Arg Glu Leu Gly Ile Asp Leu Ser Lys
1             5             10             15
Phe Lys Gly Gln Asp Glu Arg Arg Phe Leu Val Asp Ser Glu Arg Val
                20             25             30
Ser Pro Gly Gly Pro Asp Pro Gln His His
                35             40
```

(2) INFORMATION FOR SEQ ID NO:309:

- (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 903 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
(B) LOCATION: 1..903
(D) OTHER INFORMATION: / Ceres Seq. ID 1482033

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:309:

```
aatgtcgcgt gcggccacta gatttttcct gacgcggtgt ctgctccac ttccccctcct      60
ctccccagg tggcggcagc ggcggcgggg tagcatttgt gctacgaggg cttttgcaat      120
ggcggttcg gggttcggcg gcggcgaggc gttccggctc tcggccgcac caggggcccgg      180
cttactgaag ctgcacaagg gcgacatcac cctctgtgcc gtcgactgcg ccaccgacgc      240
catcgtaaat gctgctaatt agcgaatgtt aggtggcgga ggtgttgatg gagctatata      300
tcaagctgct ggaccagagc tagtgcaagc atgccggaaa gttccagagg tcaaaccagg      360
agttcgttgt cctactggag aagctaggat tactcctgct tttgagcttc ctgcctctcg      420
ggtgattcac actgtttggc ctatatatga tttggacaag catcctgagg tgtcattaaa      480
gaaggcctat gaaaatagct tgaagcttgc taaagataat ggcattcagt acatcgcaatt      540
ccctgctata tcttgtggtg tttatcgtaa tcctcccaag gaagcatcaa aaatagctgt      600
ttctaccgca cagaaatatt cagagggtat caaagaggtg cattttgttc tgttctcgga      660
tgacctttac aatatatggc gcgagactgc ccagcagttg ctatcacagt ttgagaaatg      720
aatggtccat aggcagtttg ctagcactag cagttgcccc gcagtcgttg tctagtgttg      780
agatgtgagc gccataggca gtttgccctg tgtaataaaa atgggtgtat cagacaacgt      840
ttaaactctt atgaaaccgt gtattgcacc tgtggtataa tgctgaatga gtaaagtgtt      900
gcc
```

(2) INFORMATION FOR SEQ ID NO:310:

- (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 239 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
(B) LOCATION: 1..239
(D) OTHER INFORMATION: / Ceres Seq. ID 1482034

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:310:

```
Met Ser Arg Ala Ala Thr Arg Phe Phe Leu Thr Arg Cys Leu Leu Pro
1             5             10             15
Leu Pro Leu Leu Ser Pro Arg Trp Arg Gln Arg Arg Arg Gly Ser Ile
                20             25             30
Cys Ala Thr Arg Ala Phe Ala Met Ala Ala Ser Gly Phe Gly Gly Gly
                35             40             45
```

Glu Ala Phe Arg Leu Ser Ala Ala Pro Gly Ala Gly Leu Leu Lys Leu
50 55 60
His Lys Gly Asp Ile Thr Leu Trp Ser Val Asp Cys Ala Thr Asp Ala
65 70 75 80
Ile Val Asn Ala Ala Asn Glu Arg Met Leu Gly Gly Gly Gly Val Asp
85 90 95
Gly Ala Ile His Gln Ala Ala Gly Pro Glu Leu Val Gln Ala Cys Arg
100 105 110
Lys Val Pro Glu Val Lys Pro Gly Val Arg Cys Pro Thr Gly Glu Ala
115 120 125
Arg Ile Thr Pro Ala Phe Glu Leu Pro Ala Ser Arg Val Ile His Thr
130 135 140
Val Gly Pro Ile Tyr Asp Leu Asp Lys His Pro Glu Val Ser Leu Lys
145 150 155 160
Lys Ala Tyr Glu Asn Ser Leu Lys Leu Ala Lys Asp Asn Gly Ile Gln
165 170 175
Tyr Ile Ala Phe Pro Ala Ile Ser Cys Gly Val Tyr Arg Tyr Pro Pro
180 185 190
Lys Glu Ala Ser Lys Ile Ala Val Ser Thr Ala Gln Lys Phe Ser Glu
195 200 205
Gly Ile Lys Glu Val His Phe Val Leu Phe Ser Asp Asp Leu Tyr Asn
210 215 220
Ile Trp Arg Glu Thr Ala Gln Gln Leu Leu Ser Gln Phe Glu Lys
225 230 235

(2) INFORMATION FOR SEQ ID NO:311:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 239 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..239

(D) OTHER INFORMATION: / Ceres Seq. ID 1482035

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:311:

Met Ser Arg Ala Ala Thr Arg Phe Phe Leu Thr Arg Cys Leu Leu Pro
1 5 10 15
Leu Pro Leu Leu Ser Pro Arg Trp Arg Gln Arg Arg Arg Gly Ser Ile
20 25 30
Cys Ala Thr Arg Ala Phe Ala Met Ala Ala Ser Gly Phe Gly Gly Gly
35 40 45
Glu Ala Phe Arg Leu Ser Ala Ala Pro Gly Ala Gly Leu Leu Lys Leu
50 55 60
His Lys Gly Asp Ile Thr Leu Trp Ser Val Asp Cys Ala Thr Asp Ala
65 70 75 80
Ile Val Asn Ala Ala Asn Glu Arg Met Leu Gly Gly Gly Gly Val Asp
85 90 95
Gly Ala Ile His Gln Ala Ala Gly Pro Glu Leu Val Gln Ala Cys Arg
100 105 110
Lys Val Pro Glu Val Lys Pro Gly Val Arg Cys Pro Thr Gly Glu Ala
115 120 125
Arg Ile Thr Pro Ala Phe Glu Leu Pro Ala Ser Arg Val Ile His Thr
130 135 140
Val Gly Pro Ile Tyr Asp Leu Asp Lys His Pro Glu Val Ser Leu Lys
145 150 155 160
Lys Ala Tyr Glu Asn Ser Leu Lys Leu Ala Lys Asp Asn Gly Ile Gln
165 170 175
Tyr Ile Ala Phe Pro Ala Ile Ser Cys Gly Val Tyr Arg Tyr Pro Pro

```

      180              185              190
Lys Glu Ala Ser Lys Ile Ala Val Ser Thr Ala Gln Lys Phe Ser Glu
      195              200              205
Gly Ile Lys Glu Val His Phe Val Leu Phe Ser Asp Asp Leu Tyr Asn
      210              215              220
Ile Trp Arg Glu Thr Ala Gln Gln Leu Leu Ser Gln Phe Glu Lys
      225              230              235
```

(2) INFORMATION FOR SEQ ID NO:312:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 200 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..200

(D) OTHER INFORMATION: / Ceres Seq. ID 1482036

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:312:

```

Met Ala Ala Ser Gly Phe Gly Gly Gly Glu Ala Phe Arg Leu Ser Ala
1          5          10          15
Ala Pro Gly Ala Gly Leu Leu Lys Leu His Lys Gly Asp Ile Thr Leu
      20          25          30
Trp Ser Val Asp Cys Ala Thr Asp Ala Ile Val Asn Ala Ala Asn Glu
      35          40          45
Arg Met Leu Gly Gly Gly Gly Val Asp Gly Ala Ile His Gln Ala Ala
      50          55          60
Gly Pro Glu Leu Val Gln Ala Cys Arg Lys Val Pro Glu Val Lys Pro
      65          70          75          80
Gly Val Arg Cys Pro Thr Gly Glu Ala Arg Ile Thr Pro Ala Phe Glu
      85          90          95
Leu Pro Ala Ser Arg Val Ile His Thr Val Gly Pro Ile Tyr Asp Leu
      100         105         110
Asp Lys His Pro Glu Val Ser Leu Lys Lys Ala Tyr Glu Asn Ser Leu
      115         120         125
Lys Leu Ala Lys Asp Asn Gly Ile Gln Tyr Ile Ala Phe Pro Ala Ile
      130         135         140
Ser Cys Gly Val Tyr Arg Tyr Pro Pro Lys Glu Ala Ser Lys Ile Ala
      145         150         155         160
Val Ser Thr Ala Gln Lys Phe Ser Glu Gly Ile Lys Glu Val His Phe
      165         170         175
Val Leu Phe Ser Asp Asp Leu Tyr Asn Ile Trp Arg Glu Thr Ala Gln
      180         185         190
Gln Leu Leu Ser Gln Phe Glu Lys
      195         200
```

(2) INFORMATION FOR SEQ ID NO:313:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 806 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..806

(D) OTHER INFORMATION: / Ceres Seq. ID 1482041

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:313:

```

aatttgacgc ttgttcccca cgagcttcct ctgttcacga tcgtcctcga gcttcctctg      60
ttcatcaagc tcctctgttc ttgaacatcg acgaaatcag aggctgtggc agatgcgaac      120
```



```
aaagcaattg agttggatca ttcattaatc aaagcttacc taagaaaagg gttacaactc 180
aagtgttttg agaaagaaga tggctaaaga tgtaagtgtt ttgggttttta ttttagagtt 240
ttggtcaatc agtttgctaa tgagtggcta ggttgagcat aaacgtgctt aacctttgat 300
ataacctcag tcaagcatga agaggagcta gctgaggtaa atatgaatgt ctttgtggta 360
ggctaaatat agccattgga tgtattcatt ttgtgtttgt aatatttagg ttggttaacc 420
aaattgggtg cttctaacat ggttatattg aatatgcagc ctcaagaaat tgtggcagtg 480
aaagattgac atgttttgtt tgtcttatgt gctatttatg cagctcggag atagatttat 540
ctatgaagtt gtggatgaag tgaataactt ccctcacttc tatggtccta tcaaaacctt 600
cgttcctctt cctttggatt atgttgtcaa agttgagaag ttaacattca tcaattgcaa 660
ttcacctgc agcttttttg acttgatgat tcagtggttt atgtgtaatt gcaatgtcac 720
tcttttaata atgtaattaa gagagatttg ttttctattc acaaaacagt gtatttatac 780
tattattaca atgcaagatt aagatc
```

(2) INFORMATION FOR SEQ ID NO:314:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 48 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..48
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482042

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:314:

```
Ile Cys Ser Leu Phe Pro Thr Ser Phe Leu Cys Ser Ser Ser Ser
1           5           10           15
Ser Phe Leu Cys Ser Ser Ser Ser Ser Val Leu Glu His Arg Arg Asn
20           25           30
Gln Arg Leu Trp Gln Met Arg Thr Lys Gln Leu Ser Trp Ile Ile His
35           40           45
```

(2) INFORMATION FOR SEQ ID NO:315:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 81 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..81
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482043

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:315:

```
Met Phe Cys Leu Ser Tyr Val Leu Phe Met Gln Leu Gly Asp Arg Phe
1           5           10           15
Ile Tyr Glu Val Val Asp Glu Val Asn Asn Phe Pro His Phe Tyr Gly
20           25           30
Pro Ile Lys Thr Phe Val Pro Leu Pro Leu Asp Tyr Val Val Lys Val
35           40           45
Glu Lys Leu Thr Phe Ile Asn Cys Asn Phe Thr Cys Ser Phe Phe Asp
50           55           60
Leu Met Ile Gln Trp Phe Met Cys Asn Cys Asn Val Thr Leu Leu Ile
65           70           75           80
Met
```

(2) INFORMATION FOR SEQ ID NO:316:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 72 amino acids

- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..72
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1482044
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:316:

```
Met Gln Leu Gly Asp Arg Phe Ile Tyr Glu Val Val Asp Glu Val Asn
1          5          10          15
Asn Phe Pro His Phe Tyr Gly Pro Ile Lys Thr Phe Val Pro Leu Pro
20          25          30
Leu Asp Tyr Val Val Lys Val Glu Lys Leu Thr Phe Ile Asn Cys Asn
35          40          45
Phe Thr Cys Ser Phe Phe Asp Leu Met Ile Gln Trp Phe Met Cys Asn
50          55          60
Cys Asn Val Thr Leu Leu Ile Met
65          70
```

(2) INFORMATION FOR SEQ ID NO:317:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 576 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- (ix) FEATURE:
 - (A) NAME/KEY: -
 - (B) LOCATION: 1..576
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1482045
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:317:

```
gtcggactca gtggagaaga aggaagatcc aaatcgatcc gttgaaaagg aatttcgaat      60
ttgctgtgcg atgtcgagtg cgggtggacgc tacgggaaac ccgatcccta cttcggcggt      120
tttaacggcg tcagcgaagc atataggtat gaggtgtatg ccggagaatg ttgcgttcct      180
caaatgcaag aagaatgatc caaacccaga gaagtgtctc gacaaaggtc gtgacgtcac      240
tcgctgctgt cttggcttga aaaggagatg ggattatgtt ggggtgtatgt attactacac      300
aaacgagttt gatctgtgta ggaaagagca agaagccttc gagaaagtgt gtcccttgaa      360
atgagaatca caagttcttg tcatgttttg atttgtatct cataataaag caaaatgttc      420
atTTTTgaat gagctttact ctctccatct cttgtttgtt gtcatcccat ttatttcctc      480
tcagatgctt tcgtagttag ttccaaagac aactaaatga ctcagtttta ttgttcgatg      540
gttcactaat cagcacagaa tggaacaatt gttttt
```

(2) INFORMATION FOR SEQ ID NO:318:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 120 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..120
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1482046
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:318:

```
Ser Asp Ser Val Glu Lys Lys Glu Asp Pro Asn Arg Ser Val Glu Lys
1          5          10          15
Glu Phe Arg Ile Cys Cys Ala Met Ser Ser Ala Val Asp Ala Thr Gly
20          25          30
Asn Pro Ile Pro Thr Ser Ala Val Leu Thr Ala Ser Ala Lys His Ile
35          40          45
```

Gly Met Arg Cys Met Pro Glu Asn Val Ala Phe Leu Lys Cys Lys Lys
50 55 60
Asn Asp Pro Asn Pro Glu Lys Cys Leu Asp Lys Gly Arg Asp Val Thr
65 70 75 80
Arg Cys Val Leu Gly Leu Lys Arg Arg Trp Asp Tyr Val Gly Cys Met
85 90 95
Tyr Tyr Tyr Thr Asn Glu Phe Asp Leu Cys Arg Lys Glu Gln Glu Ala
100 105 110
Phe Glu Lys Val Cys Pro Leu Lys
115 120

(2) INFORMATION FOR SEQ ID NO:319:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 97 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..97

(D) OTHER INFORMATION: / Ceres Seq. ID 1482047

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:319:

Met Ser Ser Ala Val Asp Ala Thr Gly Asn Pro Ile Pro Thr Ser Ala
1 5 10 15
Val Leu Thr Ala Ser Ala Lys His Ile Gly Met Arg Cys Met Pro Glu
20 25 30
Asn Val Ala Phe Leu Lys Cys Lys Lys Asn Asp Pro Asn Pro Glu Lys
35 40 45
Cys Leu Asp Lys Gly Arg Asp Val Thr Arg Cys Val Leu Gly Leu Lys
50 55 60
Arg Arg Trp Asp Tyr Val Gly Cys Met Tyr Tyr Thr Asn Glu Phe
65 70 75 80
Asp Leu Cys Arg Lys Glu Gln Glu Ala Phe Glu Lys Val Cys Pro Leu
85 90 95
Lys

(2) INFORMATION FOR SEQ ID NO:320:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 71 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..71

(D) OTHER INFORMATION: / Ceres Seq. ID 1482048

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:320:

Met Arg Cys Met Pro Glu Asn Val Ala Phe Leu Lys Cys Lys Lys Asn
1 5 10 15
Asp Pro Asn Pro Glu Lys Cys Leu Asp Lys Gly Arg Asp Val Thr Arg
20 25 30
Cys Val Leu Gly Leu Lys Arg Arg Trp Asp Tyr Val Gly Cys Met Tyr
35 40 45
Tyr Tyr Thr Asn Glu Phe Asp Leu Cys Arg Lys Glu Gln Glu Ala Phe
50 55 60
Glu Lys Val Cys Pro Leu Lys
65 70

(2) INFORMATION FOR SEQ ID NO:321:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 645 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..645
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482049

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:321:

aacaaagtct	cttcctttat	tcatcaatga	ctacagcaat	atcgatgaat	ccatctttgt	60
ttcgagtaat	ctgtatactc	cattcgataa	ttgcgcttac	tagtggaacc	ttaatgatgt	120
tctacacaga	gaaagcttca	atctttggac	caggaagtga	gattgctagc	aaactaaaag	180
gatcaacacc	acacgatgaa	ctactcatat	agattttctca	gtcattctct	ggtttgcttc	240
tgtttgcaat	tggtttggta	ctgttcattg	tttcgtttgt	gaaagacaaa	gagtttcata	300
gcttccttcgc	tagtgggtcc	gtgattctgt	atgtgttaat	ggctatgtgg	agggttttgt	360
tcgagtggaa	aattgaagat	cttgcttatg	aatggcctaa	acaagctctt	ggagacattg	420
ctttggctat	ttcttgggtt	ttctttcttg	tttattcttg	gagagagaag	tatgattgat	480
gtttttgatt	ttctcttttc	tttaaaaaaa	aaacttggtg	gctaagcaaa	accagatgat	540
gtattatgat	atagtttttg	atcttcagat	ttgataaaa	aggaaatgtg	aaaaagcttt	600
agattcagac	aagatcagaa	caaacaaaat	catagtttgg	gattc		

(2) INFORMATION FOR SEQ ID NO:322:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 158 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..158
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482050

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:322:

Gln	Ser	Leu	Phe	Leu	Tyr	Ser	Ser	Met	Thr	Thr	Ala	Ile	Ser	Met	Asn
1				5					10					15	
Pro	Ser	Leu	Phe	Arg	Val	Ile	Cys	Ile	Leu	His	Ser	Ile	Ile	Ala	Leu
				20				25						30	
Thr	Ser	Gly	Thr	Leu	Met	Met	Phe	Tyr	Thr	Glu	Lys	Ala	Ser	Ile	Phe
		35					40					45			
Gly	Pro	Gly	Ser	Glu	Ile	Ala	Ser	Lys	Leu	Lys	Gly	Ser	Thr	Pro	His
	50					55					60				
Asp	Glu	Leu	Leu	Ile	Gln	Ile	Ser	Gln	Ser	Phe	Ser	Gly	Leu	Leu	Leu
65				70						75				80	
Phe	Ala	Ile	Gly	Leu	Val	Leu	Phe	Met	Val	Ser	Phe	Val	Lys	Asp	Lys
				85					90					95	
Glu	Phe	His	Ser	Phe	Phe	Ala	Ser	Gly	Ser	Val	Ile	Leu	Tyr	Val	Leu
			100					105						110	
Met	Ala	Met	Trp	Arg	Val	Leu	Phe	Glu	Trp	Lys	Ile	Glu	Asp	Leu	Ala
		115						120					125		
Tyr	Glu	Trp	Pro	Lys	Gln	Ala	Leu	Gly	Asp	Ile	Ala	Leu	Ala	Ile	Ser
	130					135					140				
Trp	Val	Phe	Phe	Leu	Val	Tyr	Ser	Trp	Arg	Glu	Lys	Tyr	Asp		
145					150					155					

(2) INFORMATION FOR SEQ ID NO:323:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 150 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:

- (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..150
 (D) OTHER INFORMATION: / Ceres Seq. ID 1482051

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:323:

Met	Thr	Thr	Ala	Ile	Ser	Met	Asn	Pro	Ser	Leu	Phe	Arg	Val	Ile	Cys
1				5					10					15	
Ile	Leu	His	Ser	Ile	Ile	Ala	Leu	Thr	Ser	Gly	Thr	Leu	Met	Met	Phe
			20					25					30		
Tyr	Thr	Glu	Lys	Ala	Ser	Ile	Phe	Gly	Pro	Gly	Ser	Glu	Ile	Ala	Ser
		35				40					45				
Lys	Leu	Lys	Gly	Ser	Thr	Pro	His	Asp	Glu	Leu	Leu	Ile	Gln	Ile	Ser
	50					55					60				
Gln	Ser	Phe	Ser	Gly	Leu	Leu	Phe	Ala	Ile	Gly	Leu	Val	Leu	Phe	
65				70				75						80	
Met	Val	Ser	Phe	Val	Lys	Asp	Lys	Glu	Phe	His	Ser	Phe	Phe	Ala	Ser
				85				90						95	
Gly	Ser	Val	Ile	Leu	Tyr	Val	Leu	Met	Ala	Met	Trp	Arg	Val	Leu	Phe
			100					105					110		
Glu	Trp	Lys	Ile	Glu	Asp	Leu	Ala	Tyr	Glu	Trp	Pro	Lys	Gln	Ala	Leu
		115				120					125				
Gly	Asp	Ile	Ala	Leu	Ala	Ile	Ser	Trp	Val	Phe	Phe	Leu	Val	Tyr	Ser
	130					135					140				
Trp	Arg	Glu	Lys	Tyr	Asp										
145				150											

(2) INFORMATION FOR SEQ ID NO:324:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 144 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..144
 (D) OTHER INFORMATION: / Ceres Seq. ID 1482052

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:324:

Met	Asn	Pro	Ser	Leu	Phe	Arg	Val	Ile	Cys	Ile	Leu	His	Ser	Ile	Ile
1				5					10					15	
Ala	Leu	Thr	Ser	Gly	Thr	Leu	Met	Met	Phe	Tyr	Thr	Glu	Lys	Ala	Ser
			20					25					30		
Ile	Phe	Gly	Pro	Gly	Ser	Glu	Ile	Ala	Ser	Lys	Leu	Lys	Gly	Ser	Thr
		35				40						45			
Pro	His	Asp	Glu	Leu	Leu	Ile	Gln	Ile	Ser	Gln	Ser	Phe	Ser	Gly	Leu
	50					55					60				
Leu	Leu	Phe	Ala	Ile	Gly	Leu	Val	Leu	Phe	Met	Val	Ser	Phe	Val	Lys
65				70						75					80
Asp	Lys	Glu	Phe	His	Ser	Phe	Phe	Ala	Ser	Gly	Ser	Val	Ile	Leu	Tyr
			85					90					95		
Val	Leu	Met	Ala	Met	Trp	Arg	Val	Leu	Phe	Glu	Trp	Lys	Ile	Glu	Asp
			100					105					110		
Leu	Ala	Tyr	Glu	Trp	Pro	Lys	Gln	Ala	Leu	Gly	Asp	Ile	Ala	Leu	Ala
		115				120					125				
Ile	Ser	Trp	Val	Phe	Phe	Leu	Val	Tyr	Ser	Trp	Arg	Glu	Lys	Tyr	Asp
	130					135					140				

(2) INFORMATION FOR SEQ ID NO:325:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 623 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..623
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482053

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:325:

aagaagagat gggggaatg ggaaaggcga taggattgct gataagcggg accttggtga	60
tcaccattgc gctaatcgca acgcgactct tctctcgctc atctccgacg ttctcatcgt	120
tctcttatct tcaactcgcca ttctcgccct tctttttcgt cacctcaatg tctcgggtacc	180
tgtggatcca ttagagtggc aaatatcaca agacacagcc tgtaacattg tggcgcgctt	240
agctaatact gttggagcag ctgaatccgt tctgcgggtt gcagcaacag gacatgacaa	300
gaggctcttt gtttaagggtt tgatctgtct ttacttcttg gcagctctag gacgaatcat	360
atcggttgac cattgcctat gcaggactat gtttggtctg tctctccatg ctttttcgga	420
gttcaattag aaactccgta ttgaaccgaa gaaacggaga gattttggat tgcgaaacac	480
cttcagagtt gtaatacaca atttgccctaa acgtgttata ttctttgtcc tctttccacc	540
tttacatggt catagctttg gatagtgtga ataatgcttt cagttcctaa atgtagaaat	600
attaatcata gtttaattctt tct	

(2) INFORMATION FOR SEQ ID NO:326:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 142 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..142
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482054

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:326:

Arg Arg Asp Gly Gly Asn Gly Lys Gly Asp Arg Ile Ala Asp Lys Arg	
1 5 10 15	
Asp Leu Val Tyr His His Cys Ala Asn Arg Asn Ala Thr Leu Leu Ser	
20 25 30	
Leu Ile Ser Asp Val Leu Ile Val Leu Leu Ser Ser Leu Ala Ile Leu	
35 40 45	
Gly Leu Leu Phe Arg His Leu Asn Val Ser Val Pro Val Asp Pro Leu	
50 55 60	
Glu Trp Gln Ile Ser Gln Asp Thr Ala Cys Asn Ile Val Ala Arg Leu	
65 70 75 80	
Ala Asn Thr Val Gly Ala Ala Glu Ser Val Leu Arg Val Ala Ala Thr	
85 90 95	
Gly His Asp Lys Arg Leu Phe Val Lys Val Val Ile Cys Leu Tyr Phe	
100 105 110	
Leu Ala Ala Leu Gly Arg Ile Ile Ser Gly Asp His Cys Leu Cys Arg	
115 120 125	
Thr Met Phe Val Leu Ser Leu His Ala Phe Ser Glu Phe Asn	
130 135 140	

(2) INFORMATION FOR SEQ ID NO:327:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 505 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

Met Ala Thr Phe Val Ala Pro Ser Arg Pro Cys Ser Leu Leu Gly Arg
1 5 10 15

Arg Leu Cys Leu Pro Ser Ala Leu Leu Val Val Ser Pro Thr Asp Ala
20 25 30
Arg Ala Pro Ser Thr Ala His Gly Ala Thr Asn Leu Gln Pro Ala Ser
35 40 45
Pro Ala Ser Ser Ser Leu Arg Ser Cys Arg Pro Pro Trp Arg Thr Ser
50 55 60
Pro Ala Ala Gly Ile Phe Pro Ser Pro His Gly Val Gln Xaa Leu Ser
65 70 75 80
Ser Pro Thr Arg Val Pro Phe Pro Arg Arg Val Xaa Pro Ser Gly Val
85 90 95
Gly Gln

(2) INFORMATION FOR SEQ ID NO:330:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 542 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..542
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482069

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:330:

aaaaaagaaa aggtctaatt actcgctctt cttgctcgcc aacgccagtg nccagaggcc	60
agagcttcgt caaagacacg ccgaaaagag ggggaggcga ctcggccgag gtccggttcc	120
gaactccggt cctccgattt gcgcgtccgg atctaccagc catggcatca tcttcggacc	180
cgtggatgaa ggagtacaat gaagcatcca gacttgctga tgacatcagt tccatgattg	240
ctgatagagg gtcccttcca caatcaggcc cagaaattat gcggcatact tcagccatcc	300
ggagaaaaat aactattctt gggactagac tggatagctt ggagtcgttg cttggcagaa	360
ttcctccaaa gtcaatcact gacaaggaga tgcataagcg ccaagacatg ttttccagtt	420
tgaagtctaa agcaaagcag atggcgacaa gtttcaacat gtcaaacttt gctaacaggg	480
aggatctgct tggtcagagt aaaaaggcag atgacatgag cagagttgct gggttagata	540
ac	

(2) INFORMATION FOR SEQ ID NO:331:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 180 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..180
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482070

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:331:

Lys Arg Lys Gly Leu Ile Thr Arg Leu Ser Cys Ser Pro Thr Pro Val	
1 5 10 15	
Xaa Arg Gly Gln Ser Phe Val Lys Asp Thr Pro Lys Arg Gly Gly Gly	
20 25 30	
Asp Ser Ala Glu Val Arg Phe Arg Thr Pro Val Leu Arg Phe Ala Arg	
35 40 45	
Pro Asp Leu Pro Ala Met Ala Ser Ser Ser Asp Pro Trp Met Lys Glu	
50 55 60	
Tyr Asn Glu Ala Ser Arg Leu Ala Asp Asp Ile Ser Ser Met Ile Ala	
65 70 75 80	
Asp Arg Gly Ser Leu Pro Gln Ser Gly Pro Glu Ile Met Arg His Thr	
85 90 95	
Ser Ala Ile Arg Arg Lys Ile Thr Ile Leu Gly Thr Arg Leu Asp Ser	

100 105 110
Leu Glu Ser Leu Leu Gly Arg Ile Pro Pro Lys Ser Ile Thr Asp Lys
115 120 125
Glu Met His Lys Arg Gln Asp Met Phe Ser Ser Leu Lys Ser Lys Ala
130 135 140
Lys Gln Met Ala Thr Ser Phe Asn Met Ser Asn Phe Ala Asn Arg Glu
145 150 155 160
Asp Leu Leu Gly Gln Ser Lys Lys Ala Asp Asp Met Ser Arg Val Ala
165 170 175
Gly Leu Asp Asn
180

(2) INFORMATION FOR SEQ ID NO:332:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 127 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..127

(D) OTHER INFORMATION: / Ceres Seq. ID 1482071

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:332:

Met Ala Ser Ser Ser Asp Pro Trp Met Lys Glu Tyr Asn Glu Ala Ser
1 5 10 15
Arg Leu Ala Asp Asp Ile Ser Ser Met Ile Ala Asp Arg Gly Ser Leu
20 25 30
Pro Gln Ser Gly Pro Glu Ile Met Arg His Thr Ser Ala Ile Arg Arg
35 40 45
Lys Ile Thr Ile Leu Gly Thr Arg Leu Asp Ser Leu Glu Ser Leu Leu
50 55 60
Gly Arg Ile Pro Pro Lys Ser Ile Thr Asp Lys Glu Met His Lys Arg
65 70 75 80
Gln Asp Met Phe Ser Ser Leu Lys Ser Lys Ala Lys Gln Met Ala Thr
85 90 95
Ser Phe Asn Met Ser Asn Phe Ala Asn Arg Glu Asp Leu Leu Gly Gln
100 105 110
Ser Lys Lys Ala Asp Asp Met Ser Arg Val Ala Gly Leu Asp Asn
115 120 125

(2) INFORMATION FOR SEQ ID NO:333:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 119 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..119

(D) OTHER INFORMATION: / Ceres Seq. ID 1482072

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:333:

Met Lys Glu Tyr Asn Glu Ala Ser Arg Leu Ala Asp Asp Ile Ser Ser
1 5 10 15
Met Ile Ala Asp Arg Gly Ser Leu Pro Gln Ser Gly Pro Glu Ile Met
20 25 30
Arg His Thr Ser Ala Ile Arg Arg Lys Ile Thr Ile Leu Gly Thr Arg
35 40 45
Leu Asp Ser Ser Leu Glu Ser Leu Leu Gly Arg Ile Pro Pro Lys Ser Ile
50 55 60

Thr Asp Lys Glu Met His Lys Arg Gln Asp Met Phe Ser Ser Leu Lys
65 70 75 80
Ser Lys Ala Lys Gln Met Ala Thr Ser Phe Asn Met Ser Asn Phe Ala
85 90 95
Asn Arg Glu Asp Leu Leu Gly Gln Ser Lys Lys Ala Asp Asp Met Ser
100 105 110
Arg Val Ala Gly Leu Asp Asn
115

(2) INFORMATION FOR SEQ ID NO:334:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 652 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..652
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482073

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:334:

gaaaaacgca	accaagtcaa	ccaacgtcgg	cttgaaattc	ggccatcacc	gttcggatct	60
ttccccacc	cggttgtata	aaagcgggcg	cctgggattc	ccctctcacc	cctccttcac	120
catcagcaaa	tcggtctgcc	ctggtttccc	ccgtcgtgaa	gcagaaacct	ctctgctgcc	180
attaccgtgc	tgcgcgccgt	cgcggtgagg	cttggccaca	accgtggaac	ctgtctccat	240
atggcgtagg	cggcgtagcg	agcttcgcct	gatggatttg	cagtccagtg	ggcccataat	300
ttctcgccgg	accgcgagca	gcaacaacct	ctcctcgccg	gccatgacct	ctacgcactc	360
caagctctcc	tccgaggacc	gtcatcttcg	tgcattgtagt	cagagtaagg	cacgaggatc	420
tgaagaatca	ccctgggtatc	tggaatctca	agtgttagga	gaagagcagg	tggttcagga	480
ggagccgcct	aacactgagg	agttcgatct	gatctagggtg	gcgtttccca	gtcgacattg	540
gcgcgcagca	tccttagttc	gttttatggt	tattctttta	ttttgtaata	agtcttccgc	600
tatgtaataa	gtactctgat	gttttatgac	atttatctct	atacactctg	tg	

(2) INFORMATION FOR SEQ ID NO:335:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 68 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..68
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482074

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:335:

Lys Asn Ala Thr Lys Ser Thr Asn Val Gly Leu Lys Phe Gly His His
1 5 10 15
Arg Ser Asp Leu Ser Pro Thr Arg Leu Tyr Lys Ser Gly Arg Leu Gly
20 25 30
Phe Pro Ser His Pro Ser Phe Thr Ile Ser Lys Ser Val Cys Pro Gly
35 40 45
Phe Pro Arg Arg Glu Ala Glu Thr Ser Leu Leu Pro Leu Pro Cys Cys
50 55 60
Ala Pro Ser Arg
65

(2) INFORMATION FOR SEQ ID NO:336:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 81 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..81
 (D) OTHER INFORMATION: / Ceres Seq. ID 1482075

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:336:

```
Met Asp Leu Gln Ser Ser Gly Pro Ile Ile Ser Arg Arg Thr Ala Ser
1          5          10          15
Ser Asn Asn Leu Ser Ser Pro Ala Met Thr Ser Thr His Ser Lys Leu
20          25          30
Ser Ser Glu Asp Arg His Leu Arg Ala Cys Ser Arg Val Arg Ser Arg
35          40          45
Gly Ser Glu Glu Ser Pro Trp Tyr Leu Glu Ser Gln Val Leu Gly Glu
50          55          60
Glu Gln Val Val Gln Glu Pro Pro Asn Thr Glu Glu Phe Asp Leu
65          70          75          80
Ile
```

(2) INFORMATION FOR SEQ ID NO:337:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 57 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

- (ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..57
 (D) OTHER INFORMATION: / Ceres Seq. ID 1482076

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:337:

```
Met Thr Ser Thr His Ser Lys Leu Ser Ser Glu Asp Arg His Leu Arg
1          5          10          15
Ala Cys Ser Arg Val Arg Ser Arg Gly Ser Glu Glu Ser Pro Trp Tyr
20          25          30
Leu Glu Ser Gln Val Leu Gly Glu Glu Gln Val Val Gln Glu Glu Pro
35          40          45
Pro Asn Thr Glu Glu Phe Asp Leu Ile
50          55
```

(2) INFORMATION FOR SEQ ID NO:338:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 814 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

- (ix) FEATURE:
 (A) NAME/KEY: -
 (B) LOCATION: 1..814
 (D) OTHER INFORMATION: / Ceres Seq. ID 1482081

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:338:

```
attctacatg cgcacacttc gtcgaggaca tgagccagct acttactcgg cttcgcgcca 60
gtacagatcc ggccgacact tcatcgtcaa gctctcgcca tttacaactc cttgctccgc 120
cgacctcgac gcaaactgca acttctcgct cgcgcctgtc caggatctgc tctcgccatg 180
gacagggaag ccaagaaaga agcggttcagg aagtatcttg aatccagtg cgtgctcgat 240
accctcacga aagctcttgt ggcgctgtac gaggagaacg ataagccttc atctgcagtc 300
gaatttggtc agcagaagtt ggggtggccc tcaatctctg actatgaaaa gctcaaggca 360
gagaagctgg acttgcaatt gaagtatgat aagcttttag aaaccacaa ggaacatgc 420
agacagctgg aggaacttaa gaatatgaag tacggtgcac cctggaactg aaataacgtg 480
tggtgacact gtaaattgtat catgaagcat gtacttttta cacctctctg aagcattgct 540
```

aagctctttg tacaatggaa acatctcatg tatctgattt tagccatctg gatccctttt 600
ggattatgaa gacacccaac tcactgtagg tcccaggtat cagatatcac caatgcagga 660
taaaggatgt gacaactatc atagttgaac catgagcaat tgtttaacca gtaatccagt 720
atcgacaaag agtgtggtct attgacttga gacttctctt ggcatggctt gtaagcagat 780
tttagtagat ttcagtggaa gagatatggc gtgc

(2) INFORMATION FOR SEQ ID NO:339:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 126 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..126

(D) OTHER INFORMATION: / Ceres Seq. ID 1482082

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:339:

Phe	Tyr	Met	Pro	Thr	Leu	Arg	Arg	Gly	His	Glu	Pro	Ala	Thr	Tyr	Ser
1				5					10					15	
Ala	Ser	Arg	Gln	Tyr	Arg	Ser	Gly	Arg	His	Phe	Ile	Val	Lys	Leu	Ser
			20					25					30		
Pro	Phe	Thr	Thr	Pro	Cys	Ser	Ala	Asp	Leu	Asp	Ala	Asn	Cys	Asn	Phe
			35				40					45			
Ser	Leu	Ala	Pro	Val	Gln	Asp	Leu	Leu	Ser	Pro	Trp	Thr	Gly	Lys	Pro
	50					55				60					
Arg	Lys	Lys	Arg	Ser	Gly	Ser	Ile	Leu	Asn	Pro	Val	Ala	Cys	Ser	Ile
65					70				75					80	
Pro	Ser	Arg	Lys	Leu	Leu	Trp	Arg	Cys	Thr	Arg	Arg	Thr	Ile	Ser	Leu
				85				90					95		
His	Leu	Gln	Ser	Asn	Leu	Phe	Ser	Arg	Ser	Trp	Val	Ala	Arg	Gln	Ser
			100				105					110			
Leu	Thr	Met	Lys	Ser	Ser	Arg	Gln	Arg	Ser	Trp	Thr	Cys	Asn		
		115				120						125			

(2) INFORMATION FOR SEQ ID NO:340:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 124 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..124

(D) OTHER INFORMATION: / Ceres Seq. ID 1482083

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:340:

Met	Pro	Thr	Leu	Arg	Arg	Gly	His	Glu	Pro	Ala	Thr	Tyr	Ser	Ala	Ser
1			5					10						15	
Arg	Gln	Tyr	Arg	Ser	Gly	Arg	His	Phe	Ile	Val	Lys	Leu	Ser	Pro	Phe
			20					25				30			
Thr	Thr	Pro	Cys	Ser	Ala	Asp	Leu	Asp	Ala	Asn	Cys	Asn	Phe	Ser	Leu
			35				40					45			
Ala	Pro	Val	Gln	Asp	Leu	Leu	Ser	Pro	Trp	Thr	Gly	Lys	Pro	Arg	Lys
	50				55					60					
Lys	Arg	Ser	Gly	Ser	Ile	Leu	Asn	Pro	Val	Ala	Cys	Ser	Ile	Pro	Ser
65					70				75					80	
Arg	Lys	Leu	Leu	Trp	Arg	Cys	Thr	Arg	Arg	Thr	Ile	Ser	Leu	His	Leu
				85				90					95		
Gln	Ser	Asn	Leu	Phe	Ser	Arg	Ser	Trp	Val	Ala	Arg	Gln	Ser	Leu	Thr
			100				105					110			

Met Lys Ser Ser Arg Gln Arg Ser Trp Thr Cys Asn
115 120

(2) INFORMATION FOR SEQ ID NO:341:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 97 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..97
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482084

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:341:

Met Asp Arg Glu Ala Lys Lys Glu Ala Phe Arg Lys Tyr Leu Glu Ser
1 5 10 15
Ser Gly Val Leu Asp Thr Leu Thr Lys Ala Leu Val Ala Leu Tyr Glu
20 25 30
Glu Asn Asp Lys Pro Ser Ser Ala Val Glu Phe Val Gln Gln Lys Leu
35 40 45
Gly Gly Pro Ser Ile Ser Asp Tyr Glu Lys Leu Lys Ala Glu Lys Leu
50 55 60
Asp Leu Gln Leu Lys Tyr Asp Lys Leu Leu Glu Thr His Lys Glu Thr
65 70 75 80
Cys Arg Gln Leu Glu Leu Lys Asn Met Lys Tyr Gly Ala Pro Trp
85 90 95
Asn

(2) INFORMATION FOR SEQ ID NO:342:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 592 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..592
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482085

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:342:

gaagaatagc cttgtctagc aagaagaaga tagagggatg atgtgattat acgcaaaata 60
ctaaaaccta gggtagtagt acaagcagta gttatgagca ggccctcttct ctttctttcc 120
tgttccgttc tttttctttt tccctgcgga attcccttct tccctagtg cctcgattcg 180
atatttcgat tggattggat taccaaggga cagagggagg gaatcccaca cacacctctg 240
gccctcgcga ggccaaggga agggaagcac tcagcaccca gcagcagaag gaccgccgta 300
aatggcgctg ccggtggcga actggggacc ctggcgctgc ggacgctgtc caagcccata 360
gccagccgcc tcaagagcca ggccgctgtc caccccaagt tccgcaactt catcatcgcc 420
atcgcccagg caaaccacca gatcaccaca aagatacaga ggcgcattta tgagcatgcc 480
acagatgtgg cgatcaggcc tttggatgag cagaaagctg ttcaagctgc tacagatctc 540
atcggggaag cctttatctt ctgcgtcgtc gtttgctgct ctaatttttg ag

(2) INFORMATION FOR SEQ ID NO:343:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 97 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide

(B) LOCATION: 1..97

(D) OTHER INFORMATION: / Ceres Seq. ID 1482086

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:343:

Met Ala Leu Pro Val Ala Asn Trp Gly Pro Trp Arg Cys Gly Arg Cys
1 5 10 15
Pro Ser Pro Ser Pro Ala Ala Ser Arg Ala Arg Pro Leu Ser Thr Pro
20 25 30
Ser Ser Ala Thr Ser Ser Ser Pro Ser Pro Arg Gln Thr Thr Arg Ser
35 40 45
Pro Gln Arg Tyr Arg Gly Ala Phe Met Ser Met Pro Gln Met Trp Arg
50 55 60
Ser Gly Leu Trp Met Ser Arg Lys Leu Phe Lys Leu Leu Gln Ile Ser
65 70 75 80
Ser Gly Lys Pro Leu Ser Ser Arg Ser Leu Phe Ala Ala Leu Ile Phe
85 90 95
Glu

(2) INFORMATION FOR SEQ ID NO:344:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 624 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..624

(D) OTHER INFORMATION: / Ceres Seq. ID 1482091

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:344:

gattaaactc acagcccaac tcctcttctc gccctcgtct gacttcgttt cggacctccc 60
cagtttttcc cctccggccg ccgcacggag aagcagaagc catgcaggcc gccgccgcgc 120
gcgcccgcgc cctcctcgcc ttaccggcgg cctcggggat ccccggaata ctctccggac 180
cgatcccagg gcgcgcatca tacgccgagg gcgttctcct ttaccgtctc aatggcgctc 240
ccgcttcgcc gtcttctccg cagcatacca ggggcttctc ctctcctgc ttgcctccc 300
gatcacactg taacctccca tcgcctacca tagcttctca atggttgaat gagaaatcag 360
tacactatca catgacgaca gcacattct caacggaagc aagtdacatg gaccacccta 420
cagaagctgt agaggagatg taccagaaaa tgttgaaatc tgttgaagct gagaccatgc 480
ctccaaatgc ctggttggtg tcaatgattg atagctgctc caataaggag gacatcaaac 540
ttctttttca aattttgcag aaactcagag tatttagact atcaaattct cgcacatcgtg 600
caacttcaat gagcatctct gcag

(2) INFORMATION FOR SEQ ID NO:345:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 103 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..103

(D) OTHER INFORMATION: / Ceres Seq. ID 1482092

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:345:

Ile Lys Leu Thr Ala Gln Leu Leu Phe Ser Pro Ser Ser Asp Phe Val
1 5 10 15
Ser Asp Leu Pro Ser Phe Ser Pro Pro Ala Ala Ala Arg Arg Ser Arg
20 25 30
Ser His Ala Gly Arg Arg Arg Ala Arg Pro Pro Pro Arg Leu Thr
35 40 45
Gly Gly Leu Gly Asp Pro Arg Asn Thr Leu Arg Thr Asp Pro Arg Ala

50 55 60
Arg Ile Ile Arg Arg Gly Arg Ser Pro Leu Pro Ser Gln Trp Arg Ser
65 70 75 80
Arg Phe Ala Val Phe Ser Ala Ala Tyr Gln Gly Leu Leu Leu Leu Leu
85 90 95
Leu Arg Leu Pro Ile Thr Leu
100

(2) INFORMATION FOR SEQ ID NO:346:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 207 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..207

(D) OTHER INFORMATION: / Ceres Seq. ID 1482093

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:346:

Leu Asn Ser Gln Pro Asn Ser Ser Ser Arg Pro Arg Leu Thr Ser Phe
1 5 10 15
Arg Thr Ser Pro Val Phe Pro Leu Arg Pro Pro His Gly Glu Ala Glu
20 25 30
Ala Met Gln Ala Ala Ala Ala Arg Ala Arg Arg Leu Leu Ala Leu Pro
35 40 45
Ala Ala Ser Gly Ile Pro Gly Ile Leu Ser Gly Pro Ile Pro Gly Arg
50 55 60
Ala Ser Tyr Ala Glu Gly Val Leu Leu Tyr Arg Leu Asn Gly Ala Pro
65 70 75 80
Ala Ser Pro Ser Ser Pro Gln His Thr Arg Gly Phe Ser Ser Ser Cys
85 90 95
Phe Ala Ser Arg Ser His Cys Asn Leu Pro Ser Pro Thr Ile Ala Ser
100 105 110
Gln Trp Leu Asn Glu Lys Ser Val His Tyr His Met Thr Thr Ala His
115 120 125
Phe Ser Thr Glu Ala Ser Xaa Met Asp His Pro Thr Glu Ala Val Glu
130 135 140
Glu Met Tyr Gln Lys Met Leu Lys Ser Val Glu Ala Glu Thr Met Pro
145 150 155 160
Pro Asn Ala Trp Leu Trp Ser Met Ile Asp Ser Cys Ser Asn Lys Glu
165 170 175
Asp Ile Lys Leu Leu Phe Gln Ile Leu Gln Lys Leu Arg Val Phe Arg
180 185 190
Leu Ser Asn Leu Arg Ile Ser Ala Thr Ser Met Ser Ile Ser Ala
195 200 205

(2) INFORMATION FOR SEQ ID NO:347:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 174 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..174

(D) OTHER INFORMATION: / Ceres Seq. ID 1482094

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:347:

Met Gln Ala Ala Ala Ala Arg Ala Arg Arg Leu Leu Ala Leu Pro Ala
1 5 10 15

Ala Ser Gly Ile Pro Gly Ile Leu Ser Gly Pro Ile Pro Gly Arg Ala
20 25 30
Ser Tyr Ala Glu Gly Val Leu Leu Tyr Arg Leu Asn Gly Ala Pro Ala
35 40 45
Ser Pro Ser Ser Pro Gln His Thr Arg Gly Phe Ser Ser Ser Cys Phe
50 55 60
Ala Ser Arg Ser His Cys Asn Leu Pro Ser Pro Thr Ile Ala Ser Gln
65 70 75 80
Trp Leu Asn Glu Lys Ser Val His Tyr His Met Thr Thr Ala His Phe
85 90 95
Ser Thr Glu Ala Ser Xaa Met Asp His Pro Thr Glu Ala Val Glu Glu
100 105 110
Met Tyr Gln Lys Met Leu Lys Ser Val Glu Ala Glu Thr Met Pro Pro
115 120 125
Asn Ala Trp Leu Trp Ser Met Ile Asp Ser Cys Ser Asn Lys Glu Asp
130 135 140
Ile Lys Leu Leu Phe Gln Ile Leu Gln Lys Leu Arg Val Phe Arg Leu
145 150 155 160
Ser Asn Leu Arg Ile Ser Ala Thr Ser Met Ser Ile Ser Ala
165 170

(2) INFORMATION FOR SEQ ID NO:348:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 558 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..558
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482095

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:348:

atgtataacg cgacccccct cccaggtccc cccagcgcca aggcggcaac cgttctcccg	60
cgtcccgcac tcccgccctt ttctcttttg ctcttctctt cctcggaag cctagggctt	120
aggctttaag cgccgcgagt gtacggcggc ggcggcgcg ggcggcacta cgacggaggc	180
agcgttgcg cccgggaacg taacgcgctc ttccggcgcg ggggcttcat gccctcacag	240
tccacggtgg tcccggagaa cagcgccctc tctaagggtc ggagcgcgca gacgtgctc	300
ccgctcaccg tgaacacagc catggacgcg gcgcaaacca gcggtgacag gtctaatttc	360
gccatcaacg gcgttgaggt gtctacgatt aggcgttggt gacgcatgct aggtaagggt	420
gagcgtgtca cagatgttgc attcactctt gatgatggta ctggcaagat agatgtgaat	480
cgctgggaaa atgaggcttc cgatgctaag gagatggctg atgctaataa cgagaactat	540
gtcatagtca ttggcggt	

(2) INFORMATION FOR SEQ ID NO:349:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 110 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..110
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482096

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:349:

Met Pro Ser Gln Ser Thr Val Val Pro Glu Asn Ser Gly Leu Ser Lys
1 5 10 15
Gly Arg Ser Ala Gln Thr Leu Leu Pro Leu Thr Val Lys Gln Thr Met
20 25 30
Asp Ala Ala Gln Thr Ser Gly Asp Arg Ser Asn Phe Ala Ile Asn Gly


```

      35              40              45
Val Glu Val Ser Thr Ile Arg Leu Val Gly Arg Met Leu Gly Lys Val
      50              55              60
Glu Arg Val Thr Asp Val Val Phe Thr Leu Asp Asp Gly Thr Gly Lys
65              70              75              80
Ile Asp Val Asn Arg Trp Glu Asn Glu Ala Ser Asp Ala Lys Glu Met
      85              90              95
Ala Asp Ala Asn Asn Glu Asn Tyr Val Ile Val Ile Gly Gly
      100             105             110
```

(2) INFORMATION FOR SEQ ID NO:350:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 79 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..79

(D) OTHER INFORMATION: / Ceres Seq. ID 1482097

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:350:

```

Met Asp Ala Ala Gln Thr Ser Gly Asp Arg Ser Asn Phe Ala Ile Asn
1              5              10              15
Gly Val Glu Val Ser Thr Ile Arg Leu Val Gly Arg Met Leu Gly Lys
      20              25              30
Val Glu Arg Val Thr Asp Val Val Phe Thr Leu Asp Asp Gly Thr Gly
      35              40              45
Lys Ile Asp Val Asn Arg Trp Glu Asn Glu Ala Ser Asp Ala Lys Glu
      50              55              60
Met Ala Asp Ala Asn Asn Glu Asn Tyr Val Ile Val Ile Gly Gly
65              70              75
```

(2) INFORMATION FOR SEQ ID NO:351:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 581 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..581

(D) OTHER INFORMATION: / Ceres Seq. ID 1482102

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:351:

```

aaagatatatt gtgtagataa cagtagatta aagtctaaaa taagagagga gatggtggat      60
gaaatagggg agttttgaca gcctaattgt aattggaagc ctttcttggc ctgccctcgg      120
cgcggaaccg tcccgcactc acgcatacgc gtcgcacact cgcacgtgcc tccgtcttcg      180
ctccctcggg ccctccgcag cgtcagatcg accgtcgctc gcggasccta gcgacgccgt      240
tctcaagtcc gagccggagt agcacgagag ccttgcggna tatgagtcgc gccgcggcag      300
caagaacgat ggatgaggaa gccgagtacc tggagacggc tcgggcccgc cgctccgtgt      360
ggctcatgaa gtgccccccg gtcgtttccc gcgcctggca ggccgcctcc gcctcttcct      420
ccgatgctgc caacgccaac cccgtcgttg ccaaggctcg cctttccctt gacctgttgc      480
gccaaagaaga acgcccggaa gagcctacgc tccagttcaa gatggaattg gctcaaacta      540
acaccgggaa tacacctaag agctactctt tgaatatgtt c
```

(2) INFORMATION FOR SEQ ID NO:352:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 100 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..100

(D) OTHER INFORMATION: / Ceres Seq. ID 1482103

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:352:

```
Met Ser Arg Ala Ala Ala Arg Thr Met Asp Glu Glu Ala Glu Tyr
1          5          10          15
Leu Glu Thr Ala Arg Ala Asp Arg Ser Val Trp Leu Met Lys Cys Pro
          20          25          30
Pro Val Val Ser Arg Ala Trp Gln Ala Ala Ser Ala Ser Ser Ser Asp
          35          40          45
Ala Ala Asn Ala Asn Pro Val Val Ala Lys Val Val Leu Ser Leu Asp
          50          55          60
Leu Leu Arg Gln Glu Glu Arg Pro Glu Glu Pro Thr Leu Gln Phe Lys
65          70          75          80
Met Glu Leu Ala Gln Thr Asn Thr Gly Asn Thr Pro Lys Ser Tyr Ser
          85          90          95
Leu Asn Met Phe
          100
```

(2) INFORMATION FOR SEQ ID NO:353:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 91 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..91

(D) OTHER INFORMATION: / Ceres Seq. ID 1482104

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:353:

```
Met Asp Glu Glu Ala Glu Tyr Leu Glu Thr Ala Arg Ala Asp Arg Ser
1          5          10          15
Val Trp Leu Met Lys Cys Pro Pro Val Val Ser Arg Ala Trp Gln Ala
          20          25          30
Ala Ser Ala Ser Ser Ser Asp Ala Ala Asn Ala Asn Pro Val Val Ala
          35          40          45
Lys Val Val Leu Ser Leu Asp Leu Leu Arg Gln Glu Glu Arg Pro Glu
          50          55          60
Glu Pro Thr Leu Gln Phe Lys Met Glu Leu Ala Gln Thr Asn Thr Gly
65          70          75          80
Asn Thr Pro Lys Ser Tyr Ser Leu Asn Met Phe
          85          90
```

(2) INFORMATION FOR SEQ ID NO:354:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 72 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..72

(D) OTHER INFORMATION: / Ceres Seq. ID 1482105

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:354:

```
Met Lys Cys Pro Pro Val Val Ser Arg Ala Trp Gln Ala Ala Ser Ala
1          5          10          15
Ser Ser Ser Asp Ala Ala Asn Ala Asn Pro Val Val Ala Lys Val Val
```

20 25 30
Leu Ser Leu Asp Leu Leu Arg Gln Glu Glu Arg Pro Glu Glu Pro Thr
35 40 45
Leu Gln Phe Lys Met Glu Leu Ala Gln Thr Asn Thr Gly Asn Thr Pro
50 55 60
Lys Ser Tyr Ser Leu Asn Met Phe
65 70

(2) INFORMATION FOR SEQ ID NO:355:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 812 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..812
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482106

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:355:

gatcgagttg	gctcattaac	aattcagttt	cgtaaacaag	cctggaggga	aaaaggacac	60
atcacccaaa	ggcacagggg	atcgacccat	cgtgggaccc	gtacagcgca	ccgccagcc	120
gccggatctg	cgccggcgac	gcgcccgga	gtcgatcta	cgccgccgca	ggarggggga	180
ggcggcgctg	tgccggttct	ctgcccggtg	gcgcccgga	ccgtcccgc	gcaaggatta	240
tggatccgaa	agagaagcca	aatgtatcga	gcagtccacc	aacaccacgg	ctggactgca	300
taaaatgctt	tgatatgctc	tggttctgtt	actcaccatt	ccaccagatg	cagaattatt	360
accggtatgg	ggagttcgac	acctgcttcg	gcaagtgggg	cgatcttatg	ggctgcctcg	420
ctctcaagac	aaagcggaag	gcagaggttg	aggagatcct	catcgcgcg	gagaaggcca	480
aaccacatat	ctggacctac	cggacggctg	atgaggcatc	ggagaattgg	tggcggtatg	540
acaagcatgc	tgtgatgatg	tcaccactgc	caggttctgc	tcagcttcct	cccaggtccg	600
atgaatcttg	atagtcgagg	ggatttgtgc	aagtgttttg	tttgcgctta	tgtcacatta	660
tggcattagc	gatcatttct	gttcaaaatc	ttactgtaaa	ctacaatacc	aagagatgga	720
accattgagg	taggcagaac	atgtactgct	gaagattgag	aatttgaaat	cgccttggat	780
tcagaagcaa	ataaatgaac	gaggtttcct	tt			

(2) INFORMATION FOR SEQ ID NO:356:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 202 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..202
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482107

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:356:

Ser	Ser	Trp	Leu	Ile	Asn	Asn	Ser	Val	Ser	Leu	Thr	Ser	Leu	Glu	Gly
1			5				10				15				
Lys	Arg	Thr	His	His	Pro	Lys	Ala	Gln	Gly	Ile	Ala	Pro	Ser	Trp	Asp
		20					25				30				
Pro	Tyr	Ser	Ala	Pro	Pro	Ser	Arg	Ile	Cys	Ala	Gly	Asp	Ala	Pro	
		35					40				45				
Gly	Arg	Arg	Ile	Tyr	Ala	Ala	Ala	Gly	Xaa	Gly	Arg	Arg	Arg	Cys	Gly
		50					55				60				
Arg	Ser	Leu	Pro	Val	Ser	Ala	Arg	Ile	Arg	Pro	Ala	Ala	Arg	Ile	Met
		65					70				75				80
Asp	Pro	Lys	Glu	Lys	Pro	Asn	Val	Ser	Ser	Pro	Pro	Thr	Pro	Arg	
			85						90					95	
Leu	Asp	Cys	Ile	Lys	Cys	Phe	Asp	Met	Leu	Trp	Phe	Cys	Tyr	Ser	Pro
			100					105						110	

Phe His Gln Met Gln Asn Tyr Tyr Arg Tyr Gly Glu Phe Asp Thr Cys
115 120 125
Phe Gly Lys Trp Gly Asp Leu Met Gly Cys Leu Ala Leu Lys Thr Lys
130 135 140
Arg Lys Ala Glu Val Glu Ile Leu Ile Ala Arg Glu Lys Ala Lys
145 150 155 160
Pro His Ile Trp Thr Tyr Arg Thr Val Asp Glu Ala Ser Glu Asn Trp
165 170 175
Trp Arg Met Tyr Lys His Ala Val Met Met Ser Pro Leu Pro Gly Ser
180 185 190
Ala Gln Leu Pro Pro Arg Ser Asp Glu Ser
195 200

(2) INFORMATION FOR SEQ ID NO:357:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 123 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..123

(D) OTHER INFORMATION: / Ceres Seq. ID 1482108

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:357:

Met Asp Pro Lys Glu Lys Pro Asn Val Ser Ser Ser Pro Pro Thr Pro
1 5 10 15
Arg Leu Asp Cys Ile Lys Cys Phe Asp Met Leu Trp Phe Cys Tyr Ser
20 25 30
Pro Phe His Gln Met Gln Asn Tyr Tyr Arg Tyr Gly Glu Phe Asp Thr
35 40 45
Cys Phe Gly Lys Trp Gly Asp Leu Met Gly Cys Leu Ala Leu Lys Thr
50 55 60
Lys Arg Lys Ala Glu Val Glu Glu Ile Leu Ile Ala Arg Glu Lys Ala
65 70 75 80
Lys Pro His Ile Trp Thr Tyr Arg Thr Val Asp Glu Ala Ser Glu Asn
85 90 95
Trp Trp Arg Met Tyr Lys His Ala Val Met Met Ser Pro Leu Pro Gly
100 105 110
Ser Ala Gln Leu Pro Pro Arg Ser Asp Glu Ser
115 120

(2) INFORMATION FOR SEQ ID NO:358:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 675 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..675

(D) OTHER INFORMATION: / Ceres Seq. ID 1482113

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:358:

ataaatcccg agaccaaacc ctgcctcca ttggtcccc gccgcgcgcg ctcccagtct	60
ctacgcggaa gcagcgctc gcaccgtcc taccgaatgg cgcgacgac gaagctgtcg	120
atgagcatca agcgtgcgtc gcgtcgcac gcgtaccacc gccgtgggct ctgggccatc	180
aaggccaaga acggcggcgt cttccccaag gccgagaaac mngcmgccgc cgcggaaccc	240
aagttctacc ccgccacga cgtcaagcct cgcgttccca gcaccgcaa gcctaattcc	300
accaagctca ggtcgagcat cacgcctggg acggtgctga tcctcctcgc ggggcagaac	360
ttgggttccg cggcggcggc kgcggccggg tccgacggcg cggccgcggc gcaggcggcg	420

```
gccttccgga aggccaaacga gggcaaggcg tagctgcctg tgctgtgcat atgcatgtgt 480
ggttaattag ctggagtgtc cgggtcgctt aatctgttgg atttgatggt ttgttggttg 540
tgtgcgctg tgtttcagtg atttgctcct ttttttttct ttctcgtgga tctatcgatg 600
gatgaacatg aatgaatgaa ccgaactgca cagctccggt gtgagctgat gcatgcatgc 660
actagctagt agctg
```

(2) INFORMATION FOR SEQ ID NO:359:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 150 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..150
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482114

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:359:

```
Ile Asn Pro Glu Thr Lys Pro Ser Pro Pro Phe Val Pro Arg Arg Arg
1          5          10          15
Arg Ser Gln Ser Leu Arg Gly Ser Ser Ala Ser His Arg Ser Tyr Pro
          20          25          30
Met Ala Pro Thr Ser Lys Leu Ser Met Ser Ile Lys Arg Ala Ser Arg
          35          40          45
Ser His Ala Tyr His Arg Arg Gly Leu Trp Ala Ile Lys Ala Lys Asn
          50          55          60
Gly Gly Val Phe Pro Lys Ala Glu Lys Xaa Xaa Ala Ala Ala Glu Pro
65          70          75          80
Lys Phe Tyr Pro Ala Asp Asp Val Lys Pro Arg Val Pro Ser Thr Arg
          85          90          95
Lys Pro Asn Pro Thr Lys Leu Arg Ser Ser Ile Thr Pro Gly Thr Val
          100          105          110
Leu Ile Leu Leu Ala Gly Gln Asn Leu Gly Ser Ala Ala Ala Xaa Ala
          115          120          125
Ala Gly Ser Asp Gly Ala Ala Ala Glu Ala Ala Ala Phe Arg Lys
          130          135          140
Ala Asn Glu Gly Lys Ala
          145          150
```

(2) INFORMATION FOR SEQ ID NO:360:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 118 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..118
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482115

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:360:

```
Met Ala Pro Thr Ser Lys Leu Ser Met Ser Ile Lys Arg Ala Ser Arg
1          5          10          15
Ser His Ala Tyr His Arg Arg Gly Leu Trp Ala Ile Lys Ala Lys Asn
          20          25          30
Gly Gly Val Phe Pro Lys Ala Glu Lys Xaa Xaa Ala Ala Ala Glu Pro
          35          40          45
Lys Phe Tyr Pro Ala Asp Asp Val Lys Pro Arg Val Pro Ser Thr Arg
          50          55          60
Lys Pro Asn Pro Thr Lys Leu Arg Ser Ser Ile Thr Pro Gly Thr Val
65          70          75          80
```

Leu Ile Leu Leu Ala Gly Gln Asn Leu Gly Ser Ala Ala Ala Xaa Ala
85 90 95
Ala Gly Ser Asp Gly Ala Ala Ala Ala Gln Ala Ala Ala Phe Arg Lys
100 105 110
Ala Asn Glu Gly Lys Ala
115

(2) INFORMATION FOR SEQ ID NO:361:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 110 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..110
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482116

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:361:

Met Ser Ile Lys Arg Ala Ser Arg Ser His Ala Tyr His Arg Arg Gly
1 5 10 15
Leu Trp Ala Ile Lys Ala Lys Asn Gly Gly Val Phe Pro Lys Ala Glu
20 25 30
Lys Xaa Xaa Ala Ala Ala Glu Pro Lys Phe Tyr Pro Ala Asp Asp Val
35 40 45
Lys Pro Arg Val Pro Ser Thr Arg Lys Pro Asn Pro Thr Lys Leu Arg
50 55 60
Ser Ser Ile Thr Pro Gly Thr Val Leu Ile Leu Leu Ala Gly Gln Asn
65 70 75 80
Leu Gly Ser Ala Ala Ala Xaa Ala Ala Gly Ser Asp Gly Ala Ala Ala
85 90 95
Ala Gln Ala Ala Ala Phe Arg Lys Ala Asn Glu Gly Lys Ala
100 105 110

(2) INFORMATION FOR SEQ ID NO:362:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 871 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..871
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482117

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:362:

gagccaaaca tgcccgtccg cccagtcctc ctccaaacca aagccatcgg aggagcaact 60
ggacgaccat ggcctcgccg ctgctcaagt cacactctca gctcgccgcc gccgcccgcc 120
tgactccgt gaggagagcc gaccgctgcc ctgcgacact acacctgggc aagttccatg 180
accacgggct caggtccggc cgttctaaga gatccggttc agcgagggtg ggcgcccttc 240
cgtcgctgga cgtggtgccg ctgatggtga cgatggtgga gcacgtggac atgtcgcggg 300
actacgtcgt gaccaagtcc atctggcatc tcagcgacgt agccctcaag agcgtctata 360
ccttctacgc catgttcacc gtctggggag tctgcttctt cgcgctccatg aaggatccct 420
tctacgacag cgagacgtac aggagccagg gtggcgacgg gaccgtgcac tggactacg 480
acaggcaaga ggacctggag gcgtctgcga gggaggagct gctgcgggag gagctgctcg 540
aggagattga gcagagggtt gggggcctca gggagctgga ggaagcgagc aaggaggagc 600
agctcacaaa gtgatcacgc gcgggcgaat accgaatggg atggatacgg gctactcatc 660
agctctctat ctgagcttcg ttagcaaata agttcagact tctttactgc cctgctcaag 720
tctgtatatg gccaaaaccc aaaacgattg atcaactgcg ctactgcagt gcaatagcag 780
gatacgtata gttttttttt caagggaaac aggaggggag tgcagtgcac ccctgccccat 840
actgcatggtt attaataaaa gaaagttatt t

(2) INFORMATION FOR SEQ ID NO:363:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 203 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..203
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482118

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:363:

Ala	Lys	His	Ala	Arg	Pro	Pro	Ser	Leu	Ala	Pro	Asn	Gln	Ser	His	Arg	
1				5				10						15		
Arg	Ser	Asn	Trp	Thr	Thr	Met	Ala	Ser	Pro	Leu	Leu	Lys	Ser	His	Ser	
		20						25					30			
Gln	Leu	Ala	Ala	Ala	Ala	Ala	Leu	His	Ser	Val	Arg	Arg	Ala	Asp	Arg	
		35					40					45				
Cys	Pro	Ala	Thr	Leu	His	Leu	Gly	Lys	Phe	His	Asp	His	Gly	Leu	Arg	
	50					55				60						
Ser	Gly	Arg	Ser	Lys	Arg	Ser	Gly	Ser	Ala	Arg	Val	Gly	Ala	Phe	Pro	
65				70					75					80		
Ser	Leu	Asp	Val	Val	Pro	Leu	Met	Val	Thr	Met	Val	Glu	His	Val	Asp	
			85					90					95			
Met	Ser	Arg	Asp	Tyr	Val	Val	Thr	Lys	Ser	Ile	Trp	His	Leu	Ser	Asp	
		100						105					110			
Val	Ala	Leu	Lys	Ser	Val	Tyr	Thr	Phe	Tyr	Ala	Met	Phe	Thr	Val	Trp	
	115						120					125				
Gly	Val	Cys	Phe	Phe	Ala	Ser	Met	Lys	Asp	Pro	Phe	Tyr	Asp	Ser	Glu	
	130					135				140						
Thr	Tyr	Arg	Ser	Gln	Gly	Gly	Asp	Gly	Thr	Val	His	Trp	Tyr	Tyr	Asp	
145				150					155					160		
Arg	Gln	Glu	Asp	Leu	Glu	Ala	Ser	Ala	Arg	Glu	Glu	Leu	Leu	Arg	Glu	
			165						170					175		
Glu	Leu	Leu	Glu	Glu	Ile	Glu	Gln	Arg	Val	Gly	Gly	Leu	Arg	Glu	Leu	
		180						185					190			
Glu	Glu	Ala	Ser	Lys	Glu	Glu	Gln	Leu	Thr	Lys						
		195					200									

(2) INFORMATION FOR SEQ ID NO:364:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 181 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..181
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482119

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:364:

Met	Ala	Ser	Pro	Leu	Leu	Lys	Ser	His	Ser	Gln	Leu	Ala	Ala	Ala	Ala	
1				5					10					15		
Ala	Leu	His	Ser	Val	Arg	Arg	Ala	Asp	Arg	Cys	Pro	Ala	Thr	Leu	His	
		20						25					30			
Leu	Gly	Lys	Phe	His	Asp	His	Gly	Leu	Arg	Ser	Gly	Arg	Ser	Lys	Arg	
	35						40					45				
Ser	Gly	Ser	Ala	Arg	Val	Gly	Ala	Phe	Pro	Ser	Leu	Asp	Val	Val	Pro	
	50					55					60					
Leu	Met	Val	Thr	Met	Val	Glu	His	Val	Asp	Met	Ser	Arg	Asp	Tyr	Val	

```
65          70          75          80
Val Thr Lys Ser Ile Trp His Leu Ser Asp Val Ala Leu Lys Ser Val
          85          90          95
Tyr Thr Phe Tyr Ala Met Phe Thr Val Trp Gly Val Cys Phe Phe Ala
          100         105         110
Ser Met Lys Asp Pro Phe Tyr Asp Ser Glu Thr Tyr Arg Ser Gln Gly
          115         120         125
Gly Asp Gly Thr Val His Trp Tyr Tyr Asp Arg Gln Glu Asp Leu Glu
          130         135         140
Ala Ser Ala Arg Glu Glu Leu Leu Arg Glu Glu Leu Leu Glu Glu Ile
          145         150         155         160
Glu Gln Arg Val Gly Gly Leu Arg Glu Leu Glu Glu Ala Ser Lys Glu
          165         170         175
Glu Gln Leu Thr Lys
          180
```

(2) INFORMATION FOR SEQ ID NO:365:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 116 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..116
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482120

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:365:

```
Met Val Thr Met Val Glu His Val Asp Met Ser Arg Asp Tyr Val Val
1          5          10          15
Thr Lys Ser Ile Trp His Leu Ser Asp Val Ala Leu Lys Ser Val Tyr
          20         25         30
Thr Phe Tyr Ala Met Phe Thr Val Trp Gly Val Cys Phe Phe Ala Ser
          35         40         45
Met Lys Asp Pro Phe Tyr Asp Ser Glu Thr Tyr Arg Ser Gln Gly Gly
          50         55         60
Asp Gly Thr Val His Trp Tyr Tyr Asp Arg Gln Glu Asp Leu Glu Ala
          65         70         75         80
Ser Ala Arg Glu Glu Leu Leu Arg Glu Glu Leu Leu Glu Glu Ile Glu
          85         90         95
Gln Arg Val Gly Gly Leu Arg Glu Leu Glu Glu Ala Ser Lys Glu Glu
          100        105        110
Gln Leu Thr Lys
          115
```

(2) INFORMATION FOR SEQ ID NO:366:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 531 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..531
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482121

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:366:

```
tatggacttc attaccgcca cttataacg gatctacaga gcatccgaaa cctttaatat      60
tggctattgc gtctcaagac tctgcatcgt tttttcggga ccgtagtctt gggttcagatt      120
ctcccatatc tggattgatt gctttgtcta ctgctgttga tgctctttct cacattcatg      180
gtctaagcaa gcttaagaaa cagcttgtgt tcgctgtttt taatggtgag gcctgggggt      240
```



```
atcttggttag tcggaatttc ttacaggaat tagatgaagg cgctgcttct gtgaatggaa 300
ttagtagctt aaagattgac caggtactgg agattgggtc tggtaggaag gctatacttg 360
aggaatatcc atcattttat gtgcatgctg aagggaatcc atcagcttca aaggaaatat 420
tagatgcact gcaaagtsca gcaagtctct tggttctgat aatgttaaag taaaacaagc 480
agcttcatca aatcctgggtg ttccaccatc ttcattaatg tcattcataa g
```

(2) INFORMATION FOR SEQ ID NO:367:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 156 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..156

(D) OTHER INFORMATION: / Ceres Seq. ID 1482122

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:367:

```
Trp Thr Ser Leu Pro Pro Leu Asn Asn Gly Ser Thr Glu His Pro Lys
1          5          10          15
Pro Leu Ile Leu Ala Ile Ala Ser Gln Asp Ser Ala Ser Phe Phe Arg
20          25          30
Asp Arg Ser Leu Gly Ser Asp Ser Pro Ile Ser Gly Leu Ile Ala Leu
35          40          45
Leu Thr Ala Val Asp Ala Leu Ser His Ile His Gly Leu Ser Lys Leu
50          55          60
Lys Lys Gln Leu Val Phe Ala Val Phe Asn Gly Glu Ala Trp Gly Tyr
65          70          75          80
Leu Gly Ser Arg Lys Phe Leu Gln Glu Leu Asp Glu Gly Ala Ala Ser
85          90          95
Val Asn Gly Ile Ser Ser Leu Lys Ile Asp Gln Val Leu Glu Ile Gly
100         105         110
Ser Val Gly Lys Ala Ile Leu Glu Tyr Pro Ser Phe Tyr Val His
115         120         125
Ala Glu Gly Asn Pro Ser Ala Ser Lys Glu Ile Leu Asp Ala Leu Gln
130         135         140
Ser Xaa Ala Ser Leu Leu Val Leu Ile Met Leu Lys
145         150         155
```

(2) INFORMATION FOR SEQ ID NO:368:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 985 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -

- (B) LOCATION: 1..985

(D) OTHER INFORMATION: / Ceres Seq. ID 1482127

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:368:

```
aaatttctcg atcaaacatg ctctattgcc ctgtatctat cctctatcga agccggagac 60
cccagaaaga gtccgtaaac aactcccggc catggcgacc gcagaccggg tcgccgccac 120
cttcctctcc tctttcccca ctcaccatcc ccgccccctt tctctcggtt cctcgtgac 180
aaacctgtc ctaccctgtc cccttcgggc cgctgtcaca ggtggccac ggctcgctc 240
ccgtctccgg gtccaccgcg tcggcgccgc cgtcgcccag cttcccacca cgaatccgga 300
ggtagcttct ggggagaaga agatcagatg gtcatcaagg gctgtgcggt cttttgcgat 360
ggcagagctg gagggccgga agatgaggta ccctaccaca ggcaccgagg ggctcctcat 420
gggcattctt gttgaaggaa ctagtggcgc asaaaacttt tgcgtgctaa tggaaccaca 480
cttctcaaag tgcgtgagga ggcagcgaat gttcttggga aatcagaaat gttttacttt 540
agtcccatgc atccaccatt gacagaagct gcacaacgag cccttgattg ggctgtcaat 600
```

```
gaaaaattga aatcaggtga ggatggagaa gtaaccgcca atcatttgct actggggata 660
tggtcagata aagagtcggc tggtcataaa atcctgtatt cgcttggatt tgacgatgag 720
aaagccagtt tactggccaa aacggctggt gaagaggctg caatgagtct tagagagcaa 780
ggagagcacc taatttattc gtcaacttaa gttggtattg tgcactagct tttatgcaact 840
tcttggtgcc tcgagacggt gacctggaga ggctgcctct acaaacttta gaacttatta 900
tggagatatg ttaggtcaga tacgatattt gtactctcac gattgccgat gcctgtgaaa 960
acgttgcgct ttgtttgtca cgggg
```

(2) INFORMATION FOR SEQ ID NO:369:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 186 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..186
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482128

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:369:

```
Asn Phe Ser Ile Lys His Ala Leu Leu Pro Cys Ile Tyr Pro Leu Ser
1           5           10           15
Lys Pro Glu Thr Pro Glu Arg Val Arg Lys Gln Leu Pro Ala Met Ala
20           25           30
Thr Ala Asp Arg Val Ala Ala Thr Phe Leu Ser Ser Phe Pro Thr His
35           40           45
His Pro Arg Pro Phe Ser Ser Val Ser Leu Val Thr Asn Pro Val Leu
50           55           60
Pro Val Ser Leu Arg Ala Ala Val Thr Gly Gly Pro Arg Leu Ala Ser
65           70           75           80
Arg Leu Arg Val His Arg Val Gly Ala Ala Val Ala Gln Leu Pro Thr
85           90           95
Thr Asn Pro Glu Val Ala Ser Gly Glu Lys Lys Ile Arg Trp Ser Ser
100          105          110
Arg Ala Val Arg Ser Phe Ala Met Ala Glu Leu Glu Ala Arg Lys Met
115          120          125
Arg Tyr Pro Thr Thr Gly Thr Glu Gly Leu Leu Met Gly Ile Leu Val
130          135          140
Glu Gly Thr Ser Gly Ala Xaa Asn Phe Cys Val Leu Met Glu Pro His
145          150          155          160
Phe Ser Lys Cys Val Arg Arg Gln Arg Met Phe Leu Gly Asn Gln Lys
165          170          175
Cys Phe Thr Leu Val Pro Cys Ile His His
180          185
```

(2) INFORMATION FOR SEQ ID NO:370:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 156 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..156
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482129

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:370:

```
Met Ala Thr Ala Asp Arg Val Ala Ala Thr Phe Leu Ser Ser Phe Pro
1           5           10           15
Thr His His Pro Arg Pro Phe Ser Ser Val Ser Leu Val Thr Asn Pro
20           25           30
```

Val Leu Pro Val Ser Leu Arg Ala Ala Val Thr Gly Gly Pro Arg Leu
35 40 45
Ala Ser Arg Leu Arg Val His Arg Val Gly Ala Ala Val Ala Gln Leu
50 55 60
Pro Thr Thr Asn Pro Glu Val Ala Ser Gly Glu Lys Lys Ile Arg Trp
65 70 75 80
Ser Ser Arg Ala Val Arg Ser Phe Ala Met Ala Glu Leu Glu Ala Arg
85 90 95
Lys Met Arg Tyr Pro Thr Thr Gly Thr Glu Gly Leu Leu Met Gly Ile
100 105 110
Leu Val Glu Gly Thr Ser Gly Ala Xaa Asn Phe Cys Val Leu Met Glu
115 120 125
Pro His Phe Ser Lys Cys Val Arg Arg Gln Arg Met Phe Leu Gly Asn
130 135 140
Gln Lys Cys Phe Thr Leu Val Pro Cys Ile His His
145 150 155

(2) INFORMATION FOR SEQ ID NO:371:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 93 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..93
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482130

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:371:

Met Phe Tyr Phe Ser Pro Met His Pro Pro Leu Thr Glu Ala Ala Gln
1 5 10 15
Arg Ala Leu Asp Trp Ala Val Asn Glu Lys Leu Lys Ser Gly Glu Asp
20 25 30
Gly Glu Val Thr Ala Asn His Leu Leu Gly Ile Trp Ser Asp Lys
35 40 45
Glu Ser Ala Gly His Lys Ile Leu Tyr Ser Leu Gly Phe Asp Asp Glu
50 55 60
Lys Ala Ser Leu Leu Ala Lys Thr Ala Gly Glu Glu Ala Ala Met Ser
65 70 75 80
Leu Arg Glu Gln Gly Glu His Leu Ile Tyr Ser Ser Thr
85 90

(2) INFORMATION FOR SEQ ID NO:372:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 852 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..852
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482131

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:372:

ataggtgggt cgaacttcga aggggttcga ggacttcagc tatggcatct gtgggacgtg 60
caggacgagg agggaaagga gatggaggga agggagaagg gtcgctggct cgcagggcgt 120
ggaggcagta cctgctccag ctccagcaac atcctctccg cacaaagatg atcacggcgg 180
ggtgcctcgc cggcgtcagt gactccgtgg cgcagaagct ctctgggttc cagaagattg 240
agaaacgcag actcctgctc aagatgctct ttggttttgc atatggtggc ccatttggac 300
atttcttgca caaaattttg gattacatct tccaaggga gaaggatacc aaaaccatag 360
caaagaaggt gttattggag caagtgcacat cttctccctg gaataacata ttgttcttgt 420

```
tctattatgg atatgttggt gagaggaggg ctttgaagga ggtgacgacc aggggtgaaga 480
aacaataccc ttctgtgcaa ctccagcgctt ggatgttttg gccgatagtt ggttggataa 540
accaccagta catgccttta caattccgag tgatcttcca cagctttgtc gcatgttggt 600
gggggatttt cctgaacctt cgtgcaaggg ctatgtctct gaagcaggcc tagatggttt 660
agaaggaacg tatagcagca aagctcctgc ccggtgctaa ctaaagcagc cgaagaagga 720
ggatgctgga agctgtatcc tgcacggta caaaaaccgt tgtttatttc ctggtagtag 780
tcggtttatt tgaatgtcaa cgcattgcga gacagattat gctttttgta aaaaaaatt 840
gtgatgggag cg
```

(2) INFORMATION FOR SEQ ID NO:373:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 216 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..216

(D) OTHER INFORMATION: / Ceres Seq. ID 1482132

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:373:

```
Arg Trp Val Glu Leu Arg Arg Gly Ser Arg Thr Ser Ala Met Ala Ser
1          5          10          15
Val Gly Arg Ala Gly Arg Gly Gly Lys Gly Asp Gly Gly Lys Gly Glu
20          25          30
Gly Ser Leu Ala Arg Arg Ala Trp Arg Gln Tyr Leu Leu Gln Leu Gln
35          40          45
Gln His Pro Leu Arg Thr Lys Met Ile Thr Ala Gly Cys Leu Ala Gly
50          55          60
Val Ser Asp Ser Val Ala Gln Lys Leu Ser Gly Phe Gln Lys Ile Glu
65          70          75          80
Lys Arg Arg Leu Leu Lys Met Leu Phe Gly Phe Ala Tyr Gly Gly
85          90          95
Pro Phe Gly His Phe Leu His Lys Ile Leu Asp Tyr Ile Phe Gln Gly
100          105          110
Lys Lys Asp Thr Lys Thr Ile Ala Lys Lys Val Leu Leu Glu Gln Val
115          120          125
Thr Ser Ser Pro Trp Asn Asn Ile Leu Phe Leu Phe Tyr Tyr Gly Tyr
130          135          140
Val Val Glu Arg Arg Pro Leu Lys Glu Val Thr Thr Arg Val Lys Lys
145          150          155          160
Gln Tyr Pro Ser Val Gln Leu Ser Ala Trp Met Phe Trp Pro Ile Val
165          170          175
Gly Trp Ile Asn His Gln Tyr Met Pro Leu Gln Phe Arg Val Ile Phe
180          185          190
His Ser Phe Val Ala Cys Cys Trp Gly Ile Phe Leu Asn Leu Arg Ala
195          200          205
Arg Ala Met Ser Leu Lys Gln Ala
210          215
```

(2) INFORMATION FOR SEQ ID NO:374:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 203 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..203

(D) OTHER INFORMATION: / Ceres Seq. ID 1482133

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:374:

Met Ala Ser Val Gly Arg Ala Gly Arg Gly Gly Lys Gly Asp Gly Gly
1 5 10 15
Lys Gly Glu Gly Ser Leu Ala Arg Arg Ala Trp Arg Gln Tyr Leu Leu
20 25 30
Gln Leu Gln Gln His Pro Leu Arg Thr Lys Met Ile Thr Ala Gly Cys
35 40 45
Leu Ala Gly Val Ser Asp Ser Val Ala Gln Lys Leu Ser Gly Phe Gln
50 55 60
Lys Ile Glu Lys Arg Arg Leu Leu Leu Lys Met Leu Phe Gly Phe Ala
65 70 75 80
Tyr Gly Gly Pro Phe Gly His Phe Leu His Lys Ile Leu Asp Tyr Ile
85 90 95
Phe Gln Gly Lys Lys Asp Thr Lys Thr Ile Ala Lys Lys Val Leu Leu
100 105 110
Glu Gln Val Thr Ser Ser Pro Trp Asn Asn Ile Leu Phe Leu Phe Tyr
115 120 125
Tyr Gly Tyr Val Val Glu Arg Arg Pro Leu Lys Glu Val Thr Thr Arg
130 135 140
Val Lys Lys Gln Tyr Pro Ser Val Gln Leu Ser Ala Trp Met Phe Trp
145 150 155 160
Pro Ile Val Gly Trp Ile Asn His Gln Tyr Met Pro Leu Gln Phe Arg
165 170 175
Val Ile Phe His Ser Phe Val Ala Cys Cys Trp Gly Ile Phe Leu Asn
180 185 190
Leu Arg Ala Arg Ala Met Ser Leu Lys Gln Ala
195 200

(2) INFORMATION FOR SEQ ID NO:375:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 161 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..161
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482134

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:375:

Met Ile Thr Ala Gly Cys Leu Ala Gly Val Ser Asp Ser Val Ala Gln
1 5 10 15
Lys Leu Ser Gly Phe Gln Lys Ile Glu Lys Arg Arg Leu Leu Lys
20 25 30
Met Leu Phe Gly Phe Ala Tyr Gly Gly Pro Phe Gly His Phe Leu His
35 40 45
Lys Ile Leu Asp Tyr Ile Phe Gln Gly Lys Lys Asp Thr Lys Thr Ile
50 55 60
Ala Lys Lys Val Leu Leu Glu Gln Val Thr Ser Ser Pro Trp Asn Asn
65 70 75 80
Ile Leu Phe Leu Phe Tyr Tyr Gly Tyr Val Val Glu Arg Arg Pro Leu
85 90 95
Lys Glu Val Thr Thr Arg Val Lys Lys Gln Tyr Pro Ser Val Gln Leu
100 105 110
Ser Ala Trp Met Phe Trp Pro Ile Val Gly Trp Ile Asn His Gln Tyr
115 120 125
Met Pro Leu Gln Phe Arg Val Ile Phe His Ser Phe Val Ala Cys Cys
130 135 140
Trp Gly Ile Phe Leu Asn Leu Arg Ala Arg Ala Met Ser Leu Lys Gln
145 150 155 160

Ala

(2) INFORMATION FOR SEQ ID NO:376:

- ```
(i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 533 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
ii) MOLECULE TYPE: DNA (genomic)
ix) FEATURE:
 (A) NAME/KEY: -
 (B) LOCATION: 1..533
 (D) OTHER INFORMATION: / Ceres Seq. ID 1482135
```

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:376:

|            |            |             |            |             |            |     |
|------------|------------|-------------|------------|-------------|------------|-----|
| cgaaaaatgg | attcattcag | acgtgatcct  | tggcccttga | attggggagac | cttctgcaga | 60  |
| gacgcacaag | cccaggggga | aaaaagtttc  | agtcagtaag | ggagaataag  | cattgcaaac | 120 |
| cgaaacagag | gtaatcaata | aaggatggga  | aacatggata | aggatcaatt  | agagcatcta | 180 |
| acccagcat  | tccaaccaat | gactcgttga  | atcgtcgcgc | taggaatact  | ctgatattac | 240 |
| atctgctctg | ccattcataa | tctgttcacg  | ttctacaacg | aakcctataa  | gttgaagcaa | 300 |
| ccgcagatgc | aattgtagga | gtatttttgtt | tcttttctgt | gatttggcct  | cagcacaac  | 360 |
| agcacaaggg | catctccaaa | gacagtcagc  | taccgtgacc | gtgaggcatg  | acatcgttta | 420 |
| ttcagtgaag | gaaaaaaaaa | tgcAACgcag  | agctaaggtc | gcaagtatcc  | aagctagtca | 480 |
| tcgcTaaqct | gtaastgaqa | gtatgtatca  | gtttctacag | acgactgtgg  | aag        |     |

(2) INFORMATION FOR SEQ ID NO:377:

- ```
(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 32 amino acids
    (B) TYPE: amino acid
    (C) STRANDEDNESS:
    (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
    (A) NAME/KEY: peptide
    (B) LOCATION: 1..32
    (D) OTHER INFORMATION: / Ceres Seq. ID 1482136
```

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:377:

Arg	Lys	Met	Asp	Ser	Phe	Arg	Arg	Asp	Pro	Trp	Pro	Leu	Asn	Trp	Glu
1				5					10					15	
Thr	Phe	Cys	Arg	Asp	Ala	Gln	Ala	Gln	Gly	Glu	Lys	Ser	Phe	Ser	Gln
			20					25					30		

(2) INFORMATION FOR SEQ ID NO:378:

- ```
(i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 30 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
 (A) NAME/KEY: peptide
 (B) LOCATION: 1..30
 (D) OTHER INFORMATION: / Ceres Seq. ID 1482137
```

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:378:

Met Asp Ser Phe Arg Arg Asp Pro Trp Pro Leu Asn Trp Glu Thr Phe  
1 5 10 15  
Cys Arg Asp Ala Gln Ala Gln Gly Glu Lys Ser Phe Ser Gln  
20 25 30

(2) INFORMATION FOR SEQ ID NO:379:

- (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 450 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- (ix) FEATURE:
  - (A) NAME/KEY: -
  - (B) LOCATION: 1..450
  - (D) OTHER INFORMATION: / Ceres Seq. ID 1482142

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:379:

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| attcgaattt cgaacgccgc cggtcgctcc ctgttcctta gctctctcct ccgcggctcc  | 60  |
| gcctccggcc tccacggttt cgcaggcaga gatgaagaag gcgtccgcgc cgtcgcgcta  | 120 |
| cgcgccctac gactccccgt ccccttcgcc gcgcgcgcgc gscncttccg cggccgccgc  | 180 |
| gaccccgcca gcagcgcacg gcagcagccg cgccctgggtg gtcgcgggga gatccggccg | 240 |
| cgatctactg ggcgccaagc cgcaagccca cggcaacctt ggctccgtgc tacggcggtc  | 300 |
| catctccatg gacaagaagc cgccttcctc caagaaccag ctcccgggtc cccctgcmgc  | 360 |
| mgcmgccgca gcagcagcag cagcgaagaa caacgggtgc gggaagctgc mggggctgtm  | 420 |
| gcggaagtgt ttccagaaag cctcgtccac                                   |     |

(2) INFORMATION FOR SEQ ID NO:380:

- (i) SEQUENCE CHARACTERISTICS:
  - (A) LENGTH: 150 amino acids
  - (B) TYPE: amino acid
  - (C) STRANDEDNESS:
  - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
  - (A) NAME/KEY: peptide
  - (B) LOCATION: 1..150
  - (D) OTHER INFORMATION: / Ceres Seq. ID 1482143

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:380:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Ile | Arg | Ile | Ser | Asn | Ala | Ala | Gly | Arg | Ser | Leu | Phe | Pro | Ser | Ser | Leu |
| 1   |     |     | 5   |     |     |     |     | 10  |     |     |     |     | 15  |     |     |
| Leu | Arg | Gly | Ser | Ala | Ser | Gly | Leu | His | Gly | Phe | Ala | Gly | Arg | Asp | Glu |
|     |     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |
| Glu | Gly | Val | Arg | Gly | Val | Ala | Leu | Arg | Gly | Leu | Arg | Leu | Pro | Val | Pro |
|     |     |     | 35  |     |     |     |     | 40  |     |     |     |     | 45  |     |     |
| Phe | Ala | Ala | Pro | Arg | Xaa | Xaa | Phe | Arg | Gly | Arg | Arg | Asp | Pro | Gly | Ser |
|     |     |     | 50  |     |     |     |     | 55  |     |     |     | 60  |     |     |     |
| Ser | Ala | Arg | Gln | Gln | Pro | Arg | Pro | Gly | Gly | Arg | Gly | Glu | Ile | Arg | Pro |
|     |     |     | 65  |     |     |     |     | 70  |     |     |     | 75  |     |     | 80  |
| Arg | Ser | Thr | Gly | Arg | Gln | Ala | Ala | Ser | Pro | Arg | Gln | Pro | Arg | Leu | Arg |
|     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |     |
| Ala | Thr | Ala | Ala | His | Leu | His | Gly | Gln | Glu | Ala | Ala | Phe | Leu | Gln | Glu |
|     |     |     | 100 |     |     |     |     | 105 |     |     |     |     | 110 |     |     |
| Pro | Ala | Pro | Gly | Ser | Pro | Cys | Xaa | Xaa | Xaa | Arg | Ser | Ser | Ser | Ser | Ser |
|     |     |     | 115 |     |     |     |     | 120 |     |     |     |     | 125 |     |     |
| Glu | Glu | Gln | Arg | Trp | Arg | Glu | Ala | Xaa | Gly | Ala | Xaa | Ala | Glu | Val | Val |
|     |     |     | 130 |     |     |     |     | 135 |     |     |     |     | 140 |     |     |
| Pro | Glu | Ser | Leu | Val | His |     |     |     |     |     |     |     |     |     |     |
|     |     |     | 145 |     |     |     |     |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:381:

- (i) SEQUENCE CHARACTERISTICS:
  - (A) LENGTH: 149 amino acids
  - (B) TYPE: amino acid
  - (C) STRANDEDNESS:
  - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
  - (A) NAME/KEY: peptide

(B) LOCATION: 1..149

(D) OTHER INFORMATION: / Ceres Seq. ID 1482144

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:381:

```
Phe Glu Phe Arg Thr Pro Pro Val Ala Pro Cys Ser Leu Ala Leu Ser
1 5 10 15
Ser Ala Ala Pro Pro Ala Ser Thr Val Ser Gln Ala Glu Met Lys
 20 25 30
Lys Ala Ser Ala Ala Ser Arg Tyr Ala Ala Tyr Asp Ser Pro Ser Pro
 35 40 45
Ser Pro Arg Arg Ala Xaa Xaa Ser Ala Ala Ala Ala Thr Pro Gly Ala
 50 55 60
Ala His Gly Ser Ser Arg Ala Leu Val Val Ala Gly Arg Ser Gly Arg
65 70 75 80
Asp Leu Leu Gly Ala Lys Pro Gln Ala His Gly Asn Leu Gly Ser Val
 85 90 95
Leu Arg Arg Leu Ile Ser Met Asp Lys Lys Pro Pro Ser Ser Lys Asn
 100 105 110
Gln Leu Pro Val Pro Pro Xaa Xaa Xaa Ala Ala Ala Ala Ala Ala Ala
 115 120 125
Lys Asn Asn Gly Gly Gly Lys Leu Xaa Gly Leu Xaa Arg Lys Leu Phe
130 135 140
Gln Lys Ala Ser Ser
145
```

(2) INFORMATION FOR SEQ ID NO:382:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 119 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..119

(D) OTHER INFORMATION: / Ceres Seq. ID 1482145

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:382:

```
Met Lys Lys Ala Ser Ala Ala Ser Arg Tyr Ala Ala Tyr Asp Ser Pro
1 5 10 15
Ser Pro Ser Pro Arg Arg Ala Xaa Xaa Ser Ala Ala Ala Ala Thr Pro
 20 25 30
Gly Ala Ala His Gly Ser Ser Arg Ala Leu Val Val Ala Gly Arg Ser
 35 40 45
Gly Arg Asp Leu Leu Gly Ala Lys Pro Gln Ala His Gly Asn Leu Gly
50 55 60
Ser Val Leu Arg Arg Leu Ile Ser Met Asp Lys Lys Pro Pro Ser Ser
65 70 75 80
Lys Asn Gln Leu Pro Val Pro Pro Xaa Xaa Xaa Ala Ala Ala Ala Ala
 85 90 95
Ala Ala Lys Asn Asn Gly Gly Gly Lys Leu Xaa Gly Leu Xaa Arg Lys
100 105 110
Leu Phe Gln Lys Ala Ser Ser
115
```

(2) INFORMATION FOR SEQ ID NO:383:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 434 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:



(A) NAME/KEY: -  
(B) LOCATION: 1..434  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482153

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:383:

```
gaagggggga gcgctgcaga tagtctcaag tccgcttcct gtgctgggtcc gccgtcgctc 60
tctgttcgcc gtccgccgtg ccgtccatcc gcccgcgtcc acgcagcttg ggaataacgg 120
atcgccgctc gctatccctc gacctcggtg gaaagttcca aacgaccacg acgtcctctt 180
ggctccccgc tgtcatcagt ccgcaattcc gcatttctat tgtccttttt ccgtcgcatc 240
cgtttcgtct ctttgctcgc atctggcctc aagcccctca ggctcagct caccaccacg 300
aaccaaccga cagaaagagg gacgaatggc gagctctcag tgctgcgata acccgccggc 360
cctgaaccgg gcctgcgggg agggcaaggt cgctgcacag ttcggcgggc tcaaggccta 420
cgtygccggc cccg
```

(2) INFORMATION FOR SEQ ID NO:384:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 38 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS:  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide  
(B) LOCATION: 1..38  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482154

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:384:

```
Glu Gly Gly Ser Ala Ala Asp Ser Leu Lys Ser Ala Ser Cys Ala Gly
1 5 10 15
Pro Pro Ser Leu Ser Val Arg Arg Pro Pro Cys Arg Pro Ser Ala Pro
 20 25 30
Leu His Ala Ala Trp Glu
 35
```

(2) INFORMATION FOR SEQ ID NO:385:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 49 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS:  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide  
(B) LOCATION: 1..49  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482155

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:385:

```
Lys Gly Gly Ala Leu Gln Ile Val Ser Ser Pro Leu Pro Val Leu Val
1 5 10 15
Arg Arg Arg Ser Leu Phe Ala Val Arg Arg Ala Val His Pro Pro Arg
 20 25 30
Ser Thr Gln Leu Gly Asn Asn Gly Ser Pro Leu Ala Ile Pro Arg Pro
 35 40 45
Arg
```

(2) INFORMATION FOR SEQ ID NO:386:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 36 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS:  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..36

(D) OTHER INFORMATION: / Ceres Seq. ID 1482156

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:386:

Met Ala Ser Ser Gln Cys Cys Asp Asn Pro Pro Ala Leu Asn Pro Ala  
1 5 10 15  
Cys Gly Glu Gly Lys Val Val Asp Ser Phe Gly Gly Leu Lys Ala Tyr  
20 25 30  
Xaa Ala Gly Pro  
35

(2) INFORMATION FOR SEQ ID NO:387:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 633 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..633

(D) OTHER INFORMATION: / Ceres Seq. ID 1482157

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:387:

agagttcagg ggggagccag cgaagacaag acaagccagg ccagcggcgg aggagagggg 60  
gagagagaga gagagagcac ggcacagtag gcaggagggc gaggaggagc ttgtagaggg 120  
ttaagggaagg cgaccgccat gggggactcc agcgggtccg tgtcggtcga cgtcgagcgg 180  
atcttcttcg gcggcaagga gcatcgagta agaacgaggc atggctctct ttcggtttct 240  
gtgtatggag acgaggacaa gcccgcgctc gtaacttata cggatgtagc cttaaatcac 300  
atgtcttgct tccaaggatt gttcttctgt ccggaggctg cgtccctgtt gcttcacagt 360  
ttctgcgtgt accacatcac acctcaagga cagcagttgg gagcagctcc gatttcagct 420  
gatgtgcctg tgccatctgt cgacgacctt gcagatcagg ttgctgatgt cctcgatttt 480  
ttcagtttag ggtctgtcat gtgcttgggt gtcactgctg gtgcctatgt tctcaccctc 540  
tttgcaacta agtatcgga gagggttctt ggccctcatgt tggtttcacc tgtatgcaaa 600  
gccccctcct ggagcgagtg gctgtataat aag

(2) INFORMATION FOR SEQ ID NO:388:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 165 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..165

(D) OTHER INFORMATION: / Ceres Seq. ID 1482158

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:388:

Met Gly Asp Ser Ser Gly Ser Val Ser Val Asp Val Glu Arg Ile Phe  
1 5 10 15  
Phe Gly Gly Lys Glu His Arg Val Arg Thr Arg His Gly Ser Leu Ser  
20 25 30  
Val Ser Val Tyr Gly Asp Glu Asp Lys Pro Ala Leu Val Thr Tyr Pro  
35 40 45  
Asp Val Ala Leu Asn His Met Ser Cys Phe Gln Gly Leu Phe Phe Cys  
50 55 60  
Pro Glu Ala Ala Ser Leu Leu Leu His Ser Phe Cys Val Tyr His Ile  
65 70 75 80  
Thr Pro Gln Gly His Glu Leu Gly Ala Ala Pro Ile Ser Ala Asp Val  
85 90 95  
Pro Val Pro Ser Val Asp Asp Leu Ala Asp Gln Val Ala Asp Val Leu  
100 105 110  
Asp Phe Phe Ser Leu Gly Ser Val Met Cys Leu Gly Val Thr Ala Gly

115 120 125  
Ala Tyr Val Leu Thr Leu Phe Ala Thr Lys Tyr Arg Glu Arg Val Leu  
130 135 140  
Gly Leu Met Leu Val Ser Pro Val Cys Lys Ala Pro Ser Trp Ser Glu  
145 150 155 160  
Trp Leu Tyr Asn Lys  
165

(2) INFORMATION FOR SEQ ID NO:389:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 111 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..111
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482159

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:389:

Met Ser Cys Phe Gln Gly Leu Phe Phe Cys Pro Glu Ala Ala Ser Leu  
1 5 10 15  
Leu Leu His Ser Phe Cys Val Tyr His Ile Thr Pro Gln Gly His Glu  
20 25 30  
Leu Gly Ala Ala Pro Ile Ser Ala Asp Val Pro Val Pro Ser Val Asp  
35 40 45  
Asp Leu Ala Asp Gln Val Ala Asp Val Leu Asp Phe Phe Ser Leu Gly  
50 55 60  
Ser Val Met Cys Leu Gly Val Thr Ala Gly Ala Tyr Val Leu Thr Leu  
65 70 75 80  
Phe Ala Thr Lys Tyr Arg Glu Arg Val Leu Gly Leu Met Leu Val Ser  
85 90 95  
Pro Val Cys Lys Ala Pro Ser Trp Ser Glu Trp Leu Tyr Asn Lys  
100 105 110

(2) INFORMATION FOR SEQ ID NO:390:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 461 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..461
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482164

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:390:

ctcgagtoga gcgaaccgaa gccgaacata cccacccatc gtctcgtcgt ctcgtcgcgc 60  
gtgggctgtg ctctctctcc cccacctcc tcttttaaga cgacgccatc gccagccgac 120  
cctccctcgc cgtccggcgc cgtcctcctt cgtccttccc tctcatcaca gtttccacct 180  
cgcgaggggc tcgcgcgcgc gcccatcccg gccgatcgac tcacgaattc gcgcgcgac 240  
atattcgtgc aagggcaccc ccgcacggcc ggaagcacgg aatcacttcc ccgcccccca 300  
attccccggc tcctcggcgc mgatccctcg ccggtgttcg ctttccggcg gtttcgcgcg 360  
cgtgtcgcgc gcaggcgcag gcggctcggc tcggttgttt cctcctcgtg ccatcatcca 420  
tggaggcgaa sagcgcmcgc ggcacggcgg ggagaggagg c

(2) INFORMATION FOR SEQ ID NO:391:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 153 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..153

(D) OTHER INFORMATION: / Ceres Seq. ID 1482165

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:391:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Leu | Glu | Ser | Ser | Glu | Pro | Lys | Pro | Asn | Ile | Pro | Thr | His | Arg | Leu | Val |
| 1   |     |     |     | 5   |     |     |     | 10  |     |     |     |     |     | 15  |     |
| Val | Ser | Ser | Arg | Val | Gly | Val | Ala | Leu | Ser | Pro | Pro | Thr | Ser | Ser | Phe |
|     |     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |
| Lys | Thr | Thr | Pro | Ser | Pro | Ala | Gly | Pro | Pro | Ser | Pro | Ser | Gly | Ala | Val |
|     |     |     | 35  |     |     |     | 40  |     |     |     |     | 45  |     |     |     |
| Leu | Leu | Arg | Pro | Ser | Leu | Ser | Ser | Gln | Phe | Pro | Pro | Arg | Glu | Gly | Leu |
|     | 50  |     |     |     |     | 55  |     |     |     | 60  |     |     |     |     |     |
| Ala | Arg | Ala | Pro | Ile | Pro | Ala | Asp | Arg | Leu | Thr | Asn | Ser | Arg | Ala | Ile |
| 65  |     |     |     | 70  |     |     |     |     | 75  |     |     |     |     | 80  |     |
| Ile | Phe | Val | Gln | Gly | His | Pro | Arg | Thr | Ala | Gly | Ser | Thr | Glu | Ser | Leu |
|     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |     |
| Pro | Arg | Pro | Pro | Ile | Pro | Gly | Leu | Leu | Gly | Xaa | Asp | Pro | Ser | Pro | Val |
|     |     |     | 100 |     |     |     | 105 |     |     |     |     |     | 110 |     |     |
| Phe | Ala | Phe | Arg | Arg | Phe | Pro | Arg | Val | Ala | Gly | Arg | Arg | Arg | Arg |     |
|     |     | 115 |     |     |     | 120 |     |     |     |     | 125 |     |     |     |     |
| Leu | Gly | Ser | Val | Val | Ser | Ser | Ser | Cys | His | His | Pro | Trp | Arg | Arg | Xaa |
|     | 130 |     |     |     |     | 135 |     |     |     |     | 140 |     |     |     |     |
| Ala | Xaa | Ala | Ala | Arg | Arg | Gly | Glu | Glu |     |     |     |     |     |     |     |
| 145 |     |     |     | 150 |     |     |     |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:392:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 153 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..153

(D) OTHER INFORMATION: / Ceres Seq. ID 1482166

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:392:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Ser | Ser | Arg | Ala | Asn | Arg | Ser | Arg | Thr | Tyr | Pro | Pro | Ile | Val | Ser | Ser |
| 1   |     |     |     | 5   |     |     |     | 10  |     |     |     |     | 15  |     |     |
| Ser | Arg | Arg | Ala | Trp | Ala | Trp | Leu | Ser | Leu | Pro | Pro | Pro | Pro | Leu | Leu |
|     |     |     | 20  |     |     |     | 25  |     |     |     |     |     | 30  |     |     |
| Arg | Arg | Arg | His | Arg | Gln | Pro | Ala | Leu | Pro | Arg | Arg | Pro | Ala | Pro | Ser |
|     |     |     | 35  |     |     |     | 40  |     |     |     |     | 45  |     |     |     |
| Ser | Phe | Val | Leu | Pro | Ser | His | His | Ser | Phe | His | Leu | Ala | Arg | Gly | Ser |
|     | 50  |     |     |     |     | 55  |     |     |     | 60  |     |     |     |     |     |
| Arg | Ala | Arg | Pro | Ser | Arg | Pro | Ile | Asp | Ser | Arg | Ile | Arg | Ala | Arg | Ser |
| 65  |     |     |     | 70  |     |     |     | 75  |     |     |     |     |     | 80  |     |
| Tyr | Ser | Cys | Lys | Gly | Ile | Pro | Ala | Arg | Pro | Glu | Ala | Arg | Asn | His | Phe |
|     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |     |
| Pro | Ala | Pro | Gln | Phe | Pro | Gly | Ser | Ser | Ala | Xaa | Ile | Pro | Arg | Arg | Cys |
|     |     |     | 100 |     |     |     | 105 |     |     |     |     | 110 |     |     |     |
| Ser | Leu | Ser | Gly | Gly | Phe | Arg | Gly | Val | Ser | Arg | Ala | Gly | Ala | Gly | Gly |
|     |     | 115 |     |     |     | 120 |     |     |     |     | 125 |     |     |     |     |
| Ser | Ala | Arg | Leu | Phe | Pro | Pro | Arg | Ala | Ile | Ile | His | Gly | Gly | Glu | Xaa |
|     | 130 |     |     |     |     | 135 |     |     |     |     | 140 |     |     |     |     |
| Arg | Xaa | Arg | His | Gly | Gly | Glu | Arg | Arg |     |     |     |     |     |     |     |
| 145 |     |     |     | 150 |     |     |     |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:393:

- (i) SEQUENCE CHARACTERISTICS:  
    (A) LENGTH: 615 base pairs  
    (B) TYPE: nucleic acid  
    (C) STRANDEDNESS: single  
    (D) TOPOLOGY: linear  
(ii) MOLECULE TYPE: DNA (genomic)  
(ix) FEATURE:  
    (A) NAME/KEY: -  
    (B) LOCATION: 1..615  
    (D) OTHER INFORMATION: / Ceres Seq. ID 1482167

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:393:

|             |             |            |            |            |             |     |
|-------------|-------------|------------|------------|------------|-------------|-----|
| aaaacccaaaa | agaaggggttg | ctcccaacgc | aacgaactgc | ctttcccgtc | agcagcagca  | 60  |
| gcagctgcys  | cntgctgctg  | tccatctcca | tcctcccat  | cgcccgactg | gatttctccc  | 120 |
| tcgaattcgc  | acctccggcc  | ccccccctca | cttcgctgtg | tctcatcaac | gccggcatca  | 180 |
| ccgcgaggac  | tgggccagcg  | ctccctccct | ttctcctccc | tccgccttta | ttgctgacgg  | 240 |
| cgacgactgg  | gcgagctctg  | ccgcgcgtct | gcgctaggtg | cccaggtcct | cctcggggcac | 300 |
| ttcaccggcg  | acgagcaccc  | atcaggagcg | aaatggacga | ggctgttcct | gctttggcta  | 360 |
| ctggccaagc  | ttcaaccgac  | ggcgtgacag | agcagcctgt | gaatgtgtac | atatgggaca  | 420 |
| tggatgagac  | actcattttg  | ctcaagtcac | ttctggatgg | ctcatatgct | ggggcttttg  | 480 |
| atggcctcaa  | ggatcatgag  | aaaagtactg | aaataggaaa | gcgatgggag | aacctcattc  | 540 |
| ttgaactctg  | tgatgagcac  | ttcttttatg | aggagattga | gaactacaat | gaaccctatc  | 600 |
| tcaatgcctt  | gaatg       |            |            |            |             |     |

(2) INFORMATION FOR SEQ ID NO:394:

- (i) SEQUENCE CHARACTERISTICS:  
    (A) LENGTH: 204 amino acids  
    (B) TYPE: amino acid  
    (C) STRANDEDNESS:  
    (D) TOPOLOGY: linear  
(ii) MOLECULE TYPE: peptide  
(ix) FEATURE:  
    (A) NAME/KEY: peptide  
    (B) LOCATION: 1..204  
    (D) OTHER INFORMATION: / Ceres Seq. ID 1482168

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:394:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Asn | Gln | Lys | Glu | Gly | Leu | Leu | Pro | Thr | Gln | Arg | Thr | Ala | Phe | Pro | Val |
| 1   |     |     |     | 5   |     |     |     | 10  |     |     |     |     |     | 15  |     |
| Ser | Ser | Ser | Ser | Ser | Cys | Xaa | Xaa | Leu | Leu | Ser | Ile | Ser | Ile | Leu | Pro |
|     |     |     |     | 20  |     |     |     | 25  |     |     |     |     |     | 30  |     |
| Ile | Ala | Arg | Leu | Asp | Phe | Ser | Leu | Glu | Phe | Ala | Pro | Pro | Ala | Ser | Pro |
|     |     |     |     | 35  |     |     |     | 40  |     |     |     |     |     | 45  |     |
| Leu | Thr | Ser | Leu | Cys | Leu | Ile | Asn | Ala | Gly | Ile | Thr | Ala | Arg | Thr | Gly |
|     |     |     |     | 50  |     |     |     | 55  |     |     |     |     |     | 60  |     |
| Pro | Ala | Leu | Pro | Pro | Phe | Leu | Leu | Pro | Pro | Pro | Leu | Leu | Leu | Thr | Ala |
|     |     |     |     | 65  |     |     |     | 70  |     |     |     |     |     | 80  |     |
| Thr | Thr | Gly | Arg | Ala | Leu | Pro | Pro | Leu | Cys | Ala | Arg | Cys | Pro | Gly | Leu |
|     |     |     |     | 85  |     |     |     | 90  |     |     |     |     |     | 95  |     |
| Pro | Arg | Ala | Leu | His | Arg | Arg | Arg | Ala | Pro | Ile | Arg | Ser | Glu | Met | Asp |
|     |     |     |     | 100 |     |     |     | 105 |     |     |     |     |     | 110 |     |
| Glu | Ala | Val | Pro | Ala | Leu | Ala | Thr | Gly | Gln | Ala | Ser | Thr | Asp | Gly | Val |
|     |     |     |     | 115 |     |     |     | 120 |     |     |     |     |     | 125 |     |
| Thr | Glu | Gln | Pro | Val | Asn | Val | Tyr | Ile | Trp | Asp | Met | Asp | Glu | Thr | Leu |
|     |     |     |     | 130 |     |     |     | 135 |     |     |     |     |     | 140 |     |
| Ile | Leu | Leu | Lys | Ser | Leu | Leu | Asp | Gly | Ser | Tyr | Ala | Gly | Ala | Phe | Asp |
|     |     |     |     | 145 |     |     |     | 150 |     |     |     |     |     | 160 |     |
| Gly | Leu | Lys | Asp | His | Glu | Lys | Ser | Thr | Glu | Ile | Gly | Lys | Arg | Trp | Glu |
|     |     |     |     | 165 |     |     |     | 170 |     |     |     |     |     | 175 |     |
| Asn | Leu | Ile | Leu | Glu | Leu | Cys | Asp | Glu | His | Phe | Phe | Tyr | Glu | Glu | Ile |
|     |     |     |     | 180 |     |     |     | 185 |     |     |     |     |     | 190 |     |
| Glu | Asn | Tyr | Asn | Glu | Pro | Tyr | Leu | Asn | Ala | Leu | Asn |     |     |     |     |

195 200

(2) INFORMATION FOR SEQ ID NO:395:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 508 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..508

(D) OTHER INFORMATION: / Ceres Seq. ID 1482169

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:395:

|            |            |            |            |            |             |     |
|------------|------------|------------|------------|------------|-------------|-----|
| attcgataac | caagaacaaa | ccattgttgg | acgcgttccc | ctcctgcacg | caacctcatc  | 60  |
| tcgtcctcca | gatccaggat | ggccgtcctc | cttgagacct | tgcctccccg | agtgtctctg  | 120 |
| gtgagctact | ttgtttcgat | ttatcaggaa | atccgtttgc | ttcatgttgt | gcaggctcat  | 180 |
| atctgattgc | tggattcggc | aaaccgcgct | tggatcctgt | atcggttagt | ccttgccctgc | 240 |
| aaggttcttg | ttgtwtgtt  | ttggtggctg | agcatcgcat | gttctgcttc | tggatccaga  | 300 |
| tctggagaaa | tcgcgaagtc | gtcgtcgctt | ggttcggagc | ggatctgagg | cgacgataga  | 360 |
| tggaggcggc | gggatctctc | ggctctgcag | cctgctccac | ctcgatggat | gatgtctctg  | 420 |
| cgctgataat | tattggatca | atttcgataa | ttatwagtag | atctatgaga | tatgccgcgt  | 480 |

ggaagaggcg aggtaagctg cagcatgt

(2) INFORMATION FOR SEQ ID NO:396:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 40 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..40

(D) OTHER INFORMATION: / Ceres Seq. ID 1482170

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:396:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Phe | Asp | Asn | Gln | Glu | Gln | Thr | Ile | Val | Gly | Arg | Val | Pro | Leu | Leu | His |
| 1   |     |     | 5   |     |     |     |     |     | 10  |     |     |     |     | 15  |     |
| Ala | Thr | Ser | Ser | Arg | Pro | Pro | Asp | Pro | Gly | Trp | Pro | Ser | Ser | Leu | Arg |
|     |     |     | 20  |     |     |     |     |     | 25  |     |     |     |     | 30  |     |
| Pro | Cys | Leu | Pro | Glu | Cys | Ser | Arg |     |     |     |     |     |     |     |     |
|     |     |     | 35  |     |     |     | 40  |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:397:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 35 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..35

(D) OTHER INFORMATION: / Ceres Seq. ID 1482171

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:397:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Ala | Val | Leu | Leu | Glu | Thr | Leu | Pro | Pro | Arg | Val | Leu | Ser | Val | Ser |
| 1   |     |     | 5   |     |     |     |     |     |     | 10  |     |     |     | 15  |     |
| Tyr | Phe | Val | Ser | Ile | Tyr | Gln | Glu | Ile | Arg | Leu | Leu | His | Val | Val | Gln |
|     |     |     | 20  |     |     |     |     |     | 25  |     |     |     |     | 30  |     |
| Ala | His | Ile |     |     |     |     |     |     |     |     |     |     |     |     |     |
|     |     |     | 35  |     |     |     |     |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:398:

(i) SEQUENCE CHARACTERISTICS:  
    (A) LENGTH: 49 amino acids  
    (B) TYPE: amino acid  
    (C) STRANDEDNESS:  
    (D) TOPOLOGY: linear  
(ii) MOLECULE TYPE: peptide  
(ix) FEATURE:  
    (A) NAME/KEY: peptide  
    (B) LOCATION: 1..49  
    (D) OTHER INFORMATION: / Ceres Seq. ID 1482172  
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:398:  
Met Glu Ala Ala Gly Ser Leu Gly Leu Gln Ser Cys Ser Thr Ser Met  
1                  5                  10                  15  
Asp Asp Val Ser Ala Leu Ile Ile Ile Gly Ser Ile Ser Ile Ile Xaa  
                  20                  25                  30  
Ser Arg Ser Met Arg Tyr Ala Ala Trp Lys Arg Arg Gly Lys Leu Gln  
                  35                  40                  45  
His

(2) INFORMATION FOR SEQ ID NO:399:

(i) SEQUENCE CHARACTERISTICS:  
    (A) LENGTH: 597 base pairs  
    (B) TYPE: nucleic acid  
    (C) STRANDEDNESS: single  
    (D) TOPOLOGY: linear  
(ii) MOLECULE TYPE: DNA (genomic)  
(ix) FEATURE:  
    (A) NAME/KEY: -  
    (B) LOCATION: 1..597  
    (D) OTHER INFORMATION: / Ceres Seq. ID 1482177

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:399:

aatcctaccc cgcgggggga attctctctc agttctctcg gcgacgactg ggagaccgcc          60  
gccgccgccca tcctactoca ggtgccctga gaactcgatc ggagtcttcg ctggcgacga          120  
acaccacacca gctatcaggt gtacaaccat gtacttcacc gtgtccgccc ctgcaatgtc          180  
tctatgatcc tccagctacg gtagacgccc cgttcgctag ctgaggacct cctggttcct          240  
gtgagcaggc ggcgtgggta gctgctgcct tcaagcatgc agagacccaa cgcgccgtcg          300  
gcttgcgcta ccatcacctt cgcggagggt ctaaggaggg agatggagta ccgcaagtgg          360  
gtggagagga cccaccacaca tctgctcgtc ggaatctgcg gasccctgaa atgcagagag          420  
atttcagtgc aggaccagta cctgatgcga tcaagagaaa actagctgcc gagaccagtg          480  
tgcctccaca acaatcaagt ttcagctgtg taactggaca gaagcagccc caaaactggt          540  
acccacacaaa gaaaaagggt aaagttccac atcttccgtc gcagattctg cagtgtc

(2) INFORMATION FOR SEQ ID NO:400:

(i) SEQUENCE CHARACTERISTICS:  
    (A) LENGTH: 61 amino acids  
    (B) TYPE: amino acid  
    (C) STRANDEDNESS:  
    (D) TOPOLOGY: linear  
(ii) MOLECULE TYPE: peptide  
(ix) FEATURE:  
    (A) NAME/KEY: peptide  
    (B) LOCATION: 1..61  
    (D) OTHER INFORMATION: / Ceres Seq. ID 1482178

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:400:

Ile Leu Pro Arg Gly Gly Asn Ser Leu Ser Val Leu Ser Ala Thr Thr  
1                  5                  10                  15  
Gly Arg Pro Pro Pro Pro Ser Tyr Ser Arg Cys Pro Glu Asn Ser  
                  20                  25                  30  
Ile Gly Val Phe Ala Gly Asp Glu His Pro Pro Ala Ile Arg Cys Thr  
                  35                  40                  45

Thr Met Tyr Phe Thr Val Ser Ala Pro Ala Met Ser Leu  
50 55 60

(2) INFORMATION FOR SEQ ID NO:401:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 73 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..73
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482179

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:401:

Ser Tyr Pro Ala Gly Gly Ile Leu Ser Gln Phe Ser Arg Arg Arg Leu  
1 5 10 15  
Gly Asp Arg Arg Arg Arg His Pro Thr Pro Gly Ala Leu Arg Thr Arg  
20 25 30  
Ser Glu Ser Ser Leu Ala Thr Asn Thr His Gln Leu Ser Gly Val Gln  
35 40 45  
Pro Cys Thr Ser Pro Cys Pro Pro Leu Gln Cys Leu Tyr Asp Pro Pro  
50 55 60  
Ala Thr Val Asp Ala Pro Phe Ala Ser  
65 70

(2) INFORMATION FOR SEQ ID NO:402:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 62 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..62
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482180

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:402:

Met Gln Arg Pro Asn Ala Pro Ser Ala Cys Ala Thr Ile Thr Phe Ala  
1 5 10 15  
Glu Ala Leu Arg Arg Glu Met Glu Tyr Arg Lys Trp Val Glu Arg Thr  
20 25 30  
His Pro His Leu Leu Val Gly Ile Cys Gly Xaa Leu Lys Cys Arg Glu  
35 40 45  
Ile Ser Val Gln Asp Gln Tyr Leu Met Arg Ser Arg Glu Asn  
50 55 60

(2) INFORMATION FOR SEQ ID NO:403:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 576 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..576
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482188

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:403:

aggattcaca agtgctcgta gcaaattctac aaaatcccca accgcctctc aacaaagtct 60  
ccccacggag gtacacagct acgcgcaaac cgcgtctcgc gcgaagaatc cgcatttccc 120  
cttccccgca ccgcaccgca cccaaccccc gtcggagaga gagatggcat cgggtggcgga 180



```
gatgcagccc ctgcgcgccg cgggggtaccg cscgcgcgccg agatgaagga gaaggtggag 240
gcgtcgggtg tggacctgga ggccggggacc ggggagacgc tgtaccgccg gatctcgcgc 300
ggggagagcg ccctccgatg gggcttcgtc cgcaaggctt acggcatcct cgctgcgcas 360
tgctcctcac caccgcgcgc tccgcmetca ccgttctcca ccccaccctc aacgccacgc 420
tctccgactc cccgggmctm gcgctmgtrc tcgcmgtmmt gcccttmatc ctgatgatcc 480
cattgtatca ttatcagcac aagcaccac acaattccgt tttcctgggt ctgttcacgt 540
tggtgcttga gcttcagcat cggcgtggct tgtgct
```

(2) INFORMATION FOR SEQ ID NO:404:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 191 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..191
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482189

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:404:

```
Asp Ser Gln Val Leu Val Ala Asn Leu Gln Asn Pro Gln Pro Pro Leu
1 5 10 15
Asn Lys Val Ser Pro Arg Arg Tyr Thr Ala Thr Arg Lys Pro Arg Leu
20 25 30
Ala Arg Arg Ile Arg Ile Ser Pro Pro His Arg Thr Ala Pro Asn
35 40 45
Pro Arg Arg Arg Glu Arg Trp His Arg Trp Arg Arg Cys Ser Pro Ser
50 55 60
Arg Arg Arg Gly Thr Xaa Arg Ala Gly Asp Glu Gly Glu Gly Gly Gly
65 70 75 80
Val Gly Gly Gly Pro Gly Gly Arg Asp Arg Gly Asp Ala Val Pro Gly
85 90 95
Asp Leu Ala Arg Gly Glu Arg Pro Pro Met Gly Leu Arg Pro Gln Gly
100 105 110
Leu Arg His Pro Arg Cys Ala Xaa Leu Leu Thr Thr Ala Val Ser Xaa
115 120 125
Leu Thr Val Leu His Pro Thr Leu Asn Ala Thr Leu Ser Asp Ser Pro
130 135 140
Xaa Xaa Ala Xaa Xaa Leu Xaa Xaa Xaa Pro Xaa Ile Leu Met Ile Pro
145 150 155 160
Leu Tyr His Tyr Gln His Lys His Pro His Asn Ser Val Phe Leu Gly
165 170 175
Leu Phe Thr Leu Val Leu Glu Leu Gln His Arg Arg Gly Leu Cys
180 185 190
```

(2) INFORMATION FOR SEQ ID NO:405:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 412 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..412
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482193

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:405:

```
actccactct actccaccgc cgcacaacag atcgcacgcc tcggttcctt ccaactcgacg 60
cttctccgcg tcttaacccc tagtaccttc gctcgctctg ccgcctccgc cgacgacgcg 120
ccagatccgc gcrsagtggg gtccctccgc gcggatcgag ctcccgatcc gcgcagtggg 180
agtggcggcg agcgcaggag cgctcggccg ggggttcgcg gaggctggag acggaggagg 240
```

aagggagcgg tagttccgcg gtggttagatc cgccggtcct gtcgccggag atggactcat 300  
ctgtcgagaa gcaggggagc gtggcgctgg atccggacga gcgcrcgccg gcgtccggcg 360  
aaaccaaggc ctgcaccgag tgccacacca ccaagacccc gctctggcgc gg

(2) INFORMATION FOR SEQ ID NO:406:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 88 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..88
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482194

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:406:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Thr | Pro | Leu | Tyr | Ser | Thr | Arg | Ala | Gln | Gln | Ile | Ala | Arg | Leu | Gly | Ser |
| 1   |     |     |     | 5   |     |     |     |     | 10  |     |     |     |     | 15  |     |
| Leu | His | Ser | Thr | Leu | Leu | Pro | Leu | Leu | Thr | Pro | Ser | Thr | Phe | Ala | Arg |
|     |     |     | 20  |     |     |     | 25  |     |     |     |     |     | 30  |     |     |
| Ser | Ala | Ala | Ser | Ala | Asp | Asp | Ala | Pro | Asp | Pro | Arg | Xaa | Val | Val | Ser |
|     |     |     | 35  |     |     |     | 40  |     |     |     |     | 45  |     |     |     |
| Ser | Ala | Ala | Asp | Arg | Ala | Pro | Asp | Pro | Arg | Ser | Gly | Ser | Gly | Gly | Glu |
|     | 50  |     |     |     |     | 55  |     |     |     | 60  |     |     |     |     |     |
| Arg | Arg | Ser | Ala | Arg | Pro | Gly | Val | Pro | Arg | Gly | Trp | Arg | Arg | Arg | Arg |
| 65  |     |     |     | 70  |     |     |     |     | 75  |     |     |     |     | 80  |     |
| Lys | Gly | Ala | Val | Val | Pro | Arg | Trp |     |     |     |     |     |     |     |     |
|     |     |     |     | 85  |     |     |     |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:407:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 486 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..486
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482205

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:407:

|             |            |            |            |            |            |     |
|-------------|------------|------------|------------|------------|------------|-----|
| attcattacc  | ggaagagaaa | aaagtaactc | ggaaaagaag | gagacgccga | aaattcgaaa | 60  |
| ggggagggga  | aagcaaagct | gatggcggag | gccagggga  | aagcaaagca | aatggcggag | 120 |
| gccccgagca  | agatcgaatc | catgaggaag | tgggtcgctc | agcacaagct | ccgagccgta | 180 |
| gttgccctctg | gctaggtggg | atcagcagtt | cgatcgctta | caactggctc | cggcccaata | 240 |
| tgaagcctag  | cgtcaagatc | atccacgcaa | ggttgcatgc | tcaagctcta | accctggctg | 300 |
| cattagttgg  | ttctgcatgc | gtggagtact | atgaccagaa | gtatggttct | tctgggccaa | 360 |
| aggtggacaa  | atacacaagc | caatacctgg | cccattcgca | taaagattaa | aggtcgccat | 420 |
| gttggttcct  | gcattgccga | ttaattttgg | gtcatctctg | ggttgctcat | gagtcatgac | 480 |
| ccgcc       |            |            |            |            |            |     |

(2) INFORMATION FOR SEQ ID NO:408:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 135 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..135
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482206

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:408:

```
Ser Leu Pro Glu Glu Lys Lys Val Thr Arg Lys Arg Arg Arg Arg Arg
1 5 10 15
Lys Phe Glu Arg Gly Gly Glu Ser Lys Ala Asp Gly Gly Gly Pro Gly
20 25 30
Glu Ser Lys Ala Asn Gly Gly Gly Pro Glu Gln Asp Arg Ile His Glu
35 40 45
Glu Val Gly Arg Arg Ala Gln Ala Pro Ser Arg Ser Cys Leu Trp Leu
50 55 60
Gly Gly Ile Ser Ser Ser Ile Ala Tyr Asn Trp Ser Arg Pro Asn Met
65 70 75 80
Lys Pro Ser Val Lys Ile Ile His Ala Arg Leu His Ala Gln Ala Leu
85 90 95
Thr Leu Ala Ala Leu Val Gly Ser Ala Cys Val Glu Tyr Tyr Asp Gln
100 105 110
Lys Tyr Gly Ser Ser Gly Pro Lys Val Asp Lys Tyr Thr Ser Gln Tyr
115 120 125
Leu Ala His Ser His Lys Asp
130 135
```

(2) INFORMATION FOR SEQ ID NO:409:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 778 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..778
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482207

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:409:

```
ggcagaggca cggagcctca actccactgc cctgctgcaa ttttttctct gttagtcgat 60
cagccagcga gtgaaaccaa gaaattcatg gcgggttgaa ggagacacgg gagggaggtg 120
tgcatgttcc tggcgaggtg ctcccagcgg cgtagtcacc agtctgttga ctcatgggga 180
catggtcata gtcggcgctc gcttgctcga gtgccagcag caaccatggc cctattcgca 240
gccatagggg gtcagccttc ttgcgctctc atggcatcgg tggagctgat agagggtctt 300
ggggggctcc ccgtcgacgg gtctccagcg gccactgcag caccacgaag cacaatgttg 360
ttgcccggtt ctgccagggg ggcgctgtcg aagggtcgac ccaagaagcg gatctcgatg 420
ttgagtggcg cagcgtgctc cactttgcat ctcccagatg tggtcctcca ccggcggcga 480
gcccgggtgg aggtggagct cttgtggctg gaaggcggag gggaggaatg gatcggtaga 540
tgggagggag aggaaggtct tscggtgggg aggaatacac ggatggcgat tcgggagggg 600
acgacggcga tctactaggg tttagtttgg gcgtgagggg atgagggcgg atggcgatct 660
ggagacaatg acggcggttc agattagggt tgcgagcggc tcgatgggcg cgtacgtggg 720
gtggatccga gcggtccgcc gcgtcacaa tcaactattt tttttatgta aaacggat
```

(2) INFORMATION FOR SEQ ID NO:410:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 164 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..164
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482208

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:410:

```
Met Phe Leu Ala Arg Cys Ser Gln Arg Arg Ser His Gln Ser Val Asp
1 5 10 15
Ser Trp Gly His Gly His Ser Arg Arg Arg Leu Ala Arg Val Pro Ala
```

(2) INFORMATION FOR SEQ ID NO:411:

(A) LENGTH: 130 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS:  
(D) TOPOLOGY: linear

(ix) FEATURE:

- (A) NAME/KEY: peptide  
(B) LOCATION: 1..130  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482209

[illegible]

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 115 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS:  
(D) TOPOLOGY: linear

(ix) FEATURE:

- ```
(A) NAME/KEY: peptide
(B) LOCATION: 1..115
```

(D) OTHER INFORMATION: / Ceres Seq. ID 1482210

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:412:

```
Met Ala Ser Val Glu Leu Ile Glu Gly Leu Gly Gly Leu Pro Val Asp
1          5          10          15
Gly Ser Pro Ala Ala Thr Ala Ala Pro Arg Ser Thr Met Leu Leu Pro
20          25          30
Gly Ser Ala Arg Gly Ala Leu Ser Lys Gly Arg Pro Lys Lys Arg Ile
35          40          45
Ser Met Leu Ser Gly Ala Ala Cys Ser Thr Leu His Leu Pro Asp Val
50          55          60
Val Leu His Arg Arg Arg Ala Arg Val Glu Val Glu Leu Leu Trp Leu
65          70          75          80
Glu Gly Gly Gly Glu Trp Ile Gly Arg Trp Glu Gly Glu Glu Gly
85          90          95
Leu Xaa Val Gly Arg Asn Thr Arg Met Ala Ile Arg Glu Gly Thr Thr
100          105          110
Ala Ile Tyr
115
```

(2) INFORMATION FOR SEQ ID NO:413:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 721 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -

- (B) LOCATION: 1..721

(D) OTHER INFORMATION: / Ceres Seq. ID 1482217

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:413:

```
atcgttgggc cggcgcaaac cctagtcgcc acatcactgc ctctcacac catctgcctg      60
tggttcccat gtcattcctcc cccgtttgag gtcttcctgc tccttcagat ccgtctatgt      120
gtgtgtttgt cgtgcctgat actggctcgg aaggtattcc gatctgtttc ttcgggtgccg      180
tatatttcgt tgcgattttg gttcggttct ttcttgctct tcgtgggtcg ttgctggata      240
acacggatcg ttgatgctgt tcaagaagta ctgcgttatc ttcctgatgc aagtgttagg      300
ccctcgttac gaaggcttcc tgacgacaca aatatcttgc tgagatccaa gtgcgcaact      360
tctctctttt ttctctttct tttccgtatt tctcgccgtc tgctttttct cctctggatt      420
gaattttgcg tacagtttag tttttaccaa atgcaatcgt aacttacggg caggatggtt      480
tcagcaacga agtaaaggag gagattatct cgtcaaccgt aaggtgccgc taagagcttt      540
agaatatgaa aggcattagt ggtaacaaga tttgatttgg ttaggtgtta gtacaaaaaa      600
atgagattcg attccaatgc tgttgggggt actgctagtg aatatggccg ggcattctatc      660
acggtgttgt atgtgtacga aataatgtct gctttcgata cggtaagttt tgctttaagt      720
t
```

(2) INFORMATION FOR SEQ ID NO:414:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 153 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide

- (B) LOCATION: 1..153

(D) OTHER INFORMATION: / Ceres Seq. ID 1482218

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:414:

```
Ile Val Gly Pro Ala Gln Thr Leu Val Ala Thr Ser Leu Pro Pro His
1          5          10          15
Thr Ile Cys Leu Trp Phe Pro Cys His Pro Pro Pro Phe Glu Val Phe
20          25          30
```

Leu Leu Leu Gln Ile Arg Leu Cys Val Cys Leu Ser Cys Leu Ile Leu
35 40 45
Ala Arg Lys Val Phe Arg Ser Val Ser Ser Val Pro Tyr Ile Ser Leu
50 55 60
Arg Phe Trp Phe Gly Ser Phe Leu Leu Phe Val Gly Arg Cys Trp Ile
65 70 75 80
Thr Arg Ile Val Asp Ala Val Gln Glu Val Leu Arg Tyr Leu Pro Asp
85 90 95
Ala Ser Val Arg Pro Ser Leu Arg Arg Leu Pro Asp Asp Thr Asn Ile
100 105 110
Leu Leu Arg Ser Lys Cys Ala Thr Ser Leu Phe Phe Leu Phe Leu Phe
115 120 125
Arg Ile Ser Arg Arg Leu Leu Phe Leu Leu Trp Ile Glu Phe Cys Val
130 135 140
Gln Phe Ser Phe Tyr Gln Met Gln Ser
145 150

(2) INFORMATION FOR SEQ ID NO:415:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 883 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..883
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482219

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:415:

acgtctgttg	ctctctaccg	gagacggatc	agcgtgtcaa	ctgacagccc	tatgtccttc	60
gccgctttct	catggccgtt	tgcgcgcggg	ggcggggctg	gcagcagtgg	cgcaasaagt	120
ccgccgccac	ggcagaggag	gacgaggagc	tgggcgtgac	cccgcagctc	ctcgacttcc	180
tccggacgct	ctcgcccgcg	gccttcaagg	ccgccgcact	ccagctccaa	ggaggctcca	240
cggaggcggc	cgccggncga	cctcaccagc	tggcaggagc	ggcacgccgt	gctcgtgcta	300
tccaaagcta	aggaactcgc	caagattcgg	tatgatctgt	gccctcggca	cctgaaggat	360
aagcagttct	ggaggatata	cttcctgctc	gccaaagatt	acatctcacc	gtatgaacta	420
cgtgccatac	agaaggaaaa	actcagacgg	atggagacag	aaaactgcaa	gccccaaaca	480
gtgatctctg	ttgagggtgga	gatgcaagaa	tcgaagcgca	ctagtctctc	acaagcatca	540
gaagtagatc	tagaatctca	ggtttagttt	tgcagttata	gcttctaaca	gatctagctt	600
aggtaacgca	atcagtagcc	cttttatgat	tcctccacac	accaaatagc	tccacgagtt	660
cttcagatct	tggatcgact	ctcgtctagc	taccagtcgg	ctgtgtgctt	ttgtgtactg	720
aaaccaagta	ggtccttttc	tgcattacgc	agcatatgtg	cttggttggt	gtgctccgat	780
ccactgacat	gtaaatctag	ggtatcttgc	gcgtgaacaa	aaacgactgc	gtttcatgta	840
gctatagatt	atgtcaactt	cgattctgct	gtgcatgtgt	tgg		

(2) INFORMATION FOR SEQ ID NO:416:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 117 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..117
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482220

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:416:

Arg Leu Leu Leu Ser Thr Gly Asp Gly Ser Ala Cys Gln Leu Thr Ala
1 5 10 15
Leu Cys Pro Ser Pro Leu Ser His Gly Arg Phe Ala Ala Gly Ala Gly
20 25 30

[illegible]

(2) INFORMATION FOR SEQ ID NO:417:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 79 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..79

(D) OTHER INFORMATION: / Ceres Seq. ID 1482221

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:417:

Met	Ala	Val	Ser	Pro	Pro	Gly	Arg	Gly	Trp	Gln	Gln	Trp	Arg	Xaa	Lys
1				5					10					15	
Ser	Ala	Ala	Thr	Ala	Glu	Glu	Asp	Glu	Glu	Leu	Gly	Val	Thr	Pro	Gln
			20					25					30		
Leu	Leu	Asp	Phe	Leu	Arg	Thr	Leu	Ser	Pro	Asp	Ala	Phe	Lys	Ala	Ala
		35					40					45			
Ala	Leu	Gln	Leu	Gln	Gly	Gly	Ser	Thr	Glu	Ala	Ala	Xaa	Arg	Pro	
	50					55					60				
His	Gln	Leu	Ala	Gly	Ala	Ala	Arg	Arg	Ala	Arg	Ala	Ile	Gln	Ser	
65					70					75					

(2) INFORMATION FOR SEQ ID NO:418:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 732 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..732

(D) OTHER INFORMATION: / Ceres Seq. ID 1482230

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:418:

taaaggcattc	gacaaaatctt	ataagcagct	gacaaaaagc	catacacatc	cactcggagg	60
aacagcatat	ttaaattctag	aaaatgaaga	cgaaccgttt	cttaattggaa	tcaagtacac	120
agctatgcc	cctaccaagc	ggtttagaga	tatggaacag	ttatccggtg	gggagaagac	180
tgttcagca	ctggcttttg	tttttgccat	tcacagtttt	aggccatcac	cgtttcttc	240
attggacgaa	gtagatgctg	ctctggacaa	tttaaatgtg	gccaaagttg	ccgggtttat	300
cagatcaaaa	tcattgtgaac	gtgttggtga	tgaacaaggc	agcattggcg	agagttggtt	360
tcagagcata	gttatattct	tgaaggacag	tttctatgac	aaggccgagg	cacttgttgg	420
tgtttatagg	gactcagaac	gaagttgctc	gaggactctc	accttcgacc	tgagaaagta	480
tagggaatcg	tgaagcagct	tttgttgaat	gtttgtacta	tgtgttagt	tgccctgctc	540
atcagcttgc	tagatagctg	tcgtgagcct	tcgatgtttt	aactatctgt	atactcctag	600
tcctacataa	gtgctagctg	aacaaggacc	ctgaaatatt	catttggtag	gtggataact	660
gatgtttcga	acacgcataa	actttttttac	ctggtgtatg	aagccatttc	tccgaattac	720
tataatctgt	tt					

(2) INFORMATION FOR SEQ ID NO:419:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 163 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..163
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482231

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:419:

Lys	Gly	Ile	Asp	Lys	Ile	Tyr	Lys	Gln	Leu	Thr	Lys	Ser	His	Thr	His
1				5				10					15		
Pro	Leu	Gly	Gly	Thr	Ala	Tyr	Leu	Asn	Leu	Glu	Asn	Glu	Asp	Glu	Pro
			20					25				30			
Phe	Leu	Asn	Gly	Ile	Lys	Tyr	Thr	Ala	Met	Pro	Pro	Thr	Lys	Arg	Phe
		35					40					45			
Arg	Asp	Met	Glu	Gln	Leu	Ser	Gly	Gly	Glu	Lys	Thr	Val	Ala	Ala	Leu
	50					55					60				
Ala	Leu	Leu	Phe	Ala	Ile	His	Ser	Phe	Arg	Pro	Ser	Pro	Phe	Phe	Ile
65					70				75						80
Leu	Asp	Glu	Val	Asp	Ala	Ala	Leu	Asp	Asn	Leu	Asn	Val	Ala	Lys	Val
			85						90					95	
Ala	Gly	Phe	Ile	Arg	Ser	Lys	Ser	Cys	Glu	Arg	Val	Gly	Asp	Glu	Gln
			100					105					110		
Gly	Ser	Asp	Gly	Glu	Ser	Gly	Phe	Gln	Ser	Ile	Val	Ile	Ser	Leu	Lys
		115					120					125			
Asp	Ser	Phe	Tyr	Asp	Lys	Ala	Glu	Ala	Leu	Val	Gly	Val	Tyr	Arg	Asp
	130					135					140				
Ser	Glu	Arg	Ser	Cys	Ser	Arg	Thr	Leu	Thr	Phe	Asp	Leu	Arg	Lys	Tyr
145					150					155					160
Arg	Glu	Ser													

(2) INFORMATION FOR SEQ ID NO:420:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 122 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..122
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482232

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:420:

Met	Pro	Pro	Thr	Lys	Arg	Phe	Arg	Asp	Met	Glu	Gln	Leu	Ser	Gly	Gly
1				5					10					15	
Glu	Lys	Thr	Val	Ala	Ala	Leu	Ala	Leu	Phe	Ala	Ile	His	Ser	Phe	
			20					25				30			
Arg	Pro	Ser	Pro	Phe	Phe	Ile	Leu	Asp	Glu	Val	Asp	Ala	Ala	Leu	Asp
		35					40					45			
Asn	Leu	Asn	Val	Ala	Lys	Val	Ala	Gly	Phe	Ile	Arg	Ser	Lys	Ser	Cys
	50					55					60				
Glu	Arg	Val	Gly	Asp	Glu	Gln	Gly	Ser	Asp	Gly	Glu	Ser	Gly	Phe	Gln
65					70					75					80
Ser	Ile	Val	Ile	Ser	Leu	Lys	Asp	Ser	Phe	Tyr	Asp	Lys	Ala	Glu	Ala
			85						90					95	
Leu	Val	Gly	Val	Tyr	Arg	Asp	Ser	Glu	Arg	Ser	Cys	Ser	Arg	Thr	Leu

100 105 110
Thr Phe Asp Leu Arg Lys Tyr Arg Glu Ser
115 120
(2) INFORMATION FOR SEQ ID NO:421:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 113 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
(A) NAME/KEY: peptide
(B) LOCATION: 1..113
(D) OTHER INFORMATION: / Ceres Seq. ID 1482233
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:421:
Met Glu Gln Leu Ser Gly Gly Glu Lys Thr Val Ala Ala Leu Ala Leu
1 5 10 15
Leu Phe Ala Ile His Ser Phe Arg Pro Ser Pro Phe Phe Ile Leu Asp
20 25 30
Glu Val Asp Ala Ala Leu Asp Asn Leu Asn Val Ala Lys Val Ala Gly
35 40 45
Phe Ile Arg Ser Lys Ser Cys Glu Arg Val Gly Asp Glu Gln Gly Ser
50 55 60
Asp Gly Glu Ser Gly Phe Gln Ser Ile Val Ile Ser Leu Lys Asp Ser
65 70 75 80
Phe Tyr Asp Lys Ala Glu Ala Leu Val Gly Val Tyr Arg Asp Ser Glu
85 90 95
Arg Ser Cys Ser Arg Thr Leu Thr Phe Asp Leu Arg Lys Tyr Arg Glu
100 105 110
Ser

(2) INFORMATION FOR SEQ ID NO:422:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 773 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: DNA (genomic)
(ix) FEATURE:
(A) NAME/KEY: -
(B) LOCATION: 1..773
(D) OTHER INFORMATION: / Ceres Seq. ID 1482234
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:422:
ctccgccgcc aggcacgacgg caaatvcgcc cagacagggt ggacgtcgac gccggcggaat 60
cctgctcgga ttgcacacat caccgccacc cgtcgtgcgc cgacatctgt aggtcgccag 120
ccaacaacct tagactgagg caccctgaat ccatctgcta ttgttcagct tgggtgttcgg 180
gcaatccttg ttctcgccctc agcacaaata gatcgccaag atgaatagaa gttggttgaa 240
tggtacattg ttttcccctg aatatatcaa tgggtgcaaa gaatttatga gctttattca 300
aagaaaattc ggtgaggatg aagatatatt gtgtccatgt agtagatgtc tcaaccaaaa 360
gtcctttcat caagcctttg tggagaagca tatattaatg aatgggatgg aaagtacata 420
tactcgatgg attcatcatg gagagaactt tgaggaagat gccgggtcatt cgatacatgg 480
gacagggttg attgatgatg acagctatgg tgatgattgt tttgatggga tgttacaaga 540
cctatgcact gcataagagc aagataaaga ggatggtgaa aatgaggatg gagacaatac 600
taatgatgac aatgagtcac tttatagtgt rgtgctgaaa gaggcgaaac gtcataattta 660
tcttggttgt accaaatttt caaggtrrtc ctttgwtgta aagcttcttc atatgaagtc 720
attatatagg atcactaact ctgcatktac tgcarkatww aagttgttgg ttg
(2) INFORMATION FOR SEQ ID NO:423:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 73 amino acids

- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..73
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1482235
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:423:

```
Pro Pro Pro Gly Arg Arg Gln Xaa Arg Pro Asp Arg Val Asp Val Asp
1          5          10          15
Ala Gly Glu Ser Cys Ser Asp Cys Thr His His Arg His Pro Ser Cys
          20          25          30
Ala Asp Ile Cys Arg Ser Pro Ala Asn Asn Leu Arg Leu Arg His Pro
          35          40          45
Glu Ser Ile Cys Tyr Cys Ser Ala Trp Cys Ser Gly Asn Pro Cys Ser
          50          55          60
Arg Leu Ser Thr Asn Arg Ser Pro Arg
65          70
```

(2) INFORMATION FOR SEQ ID NO:424:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 111 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..111
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1482236
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:424:

```
Met Asn Arg Ser Trp Leu Asn Gly Thr Leu Phe Ser Pro Glu Tyr Ile
1          5          10          15
Asn Gly Val Lys Glu Phe Met Ser Phe Ile Gln Arg Lys Phe Gly Glu
          20          25          30
Asp Glu Asp Ile Leu Cys Pro Cys Ser Arg Cys Leu Asn Gln Lys Ser
          35          40          45
Phe His Gln Ala Phe Val Glu Lys His Ile Leu Met Asn Gly Met Glu
          50          55          60
Ser Thr Tyr Thr Arg Trp Ile His His Gly Glu Asn Phe Glu Glu Asp
65          70          75          80
Ala Gly His Ser Ile His Gly Thr Gly Val Ile Asp Asp Asp Ser Tyr
          85          90          95
Gly Asp Asp Cys Phe Asp Gly Met Leu Gln Asp Leu Cys Thr Ala
          100          105          110
```

(2) INFORMATION FOR SEQ ID NO:425:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 89 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..89
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1482237
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:425:

```
Met Ser Phe Ile Gln Arg Lys Phe Gly Glu Asp Glu Asp Ile Leu Cys
1          5          10          15
```

Pro Cys Ser Arg Cys Leu Asn Gln Lys Ser Phe His Gln Ala Phe Val
20 25 30
Glu Lys His Ile Leu Met Asn Gly Met Glu Ser Thr Tyr Thr Arg Trp
35 40 45
Ile His His Gly Glu Asn Phe Glu Glu Asp Ala Gly His Ser Ile His
50 55 60
Gly Thr Gly Val Ile Asp Asp Ser Tyr Gly Asp Asp Cys Phe Asp
65 70 75 80
Gly Met Leu Gln Asp Leu Cys Thr Ala
85

(2) INFORMATION FOR SEQ ID NO:426:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 501 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..501
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482238

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:426:

cacggcgccc gtcctctgtc gttgaaggac agggagcggc ggctagggtt tcgcgggtgct	60
gtggcgcccg acgcccgtcc tctactatcc gtggcgacca tcgtcggcta tccgcggast	120
gtggcgatcg gcctgtgctc ctatccgcgg ttgcccagga aagtactatg ttgttgatgc	180
tggatatcca aatagggatg agtacttgcc cccgtacaaa ggacaactgt atcatgttcc	240
ggaatggaga aatgatcctc cacctaattg ctcactcgaa ggtgaagcat gggaagtgar	300
tcacaggtcc aacgacctcc atgaaggtaa agcacatggc ttcaagtkag tcgcaagtcc	360
aatcgagtcc atgaaggta agcatgggag gtccaagtra atctggaaag aataacggtg	420
gaagtaggtt gggccttata ataggggagg agtagtagaa attattttcc gcgtagtctg	480
ggttttaatt atttagataa g	

(2) INFORMATION FOR SEQ ID NO:427:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 58 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..58
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482239

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:427:

His Gly Ala Arg Pro Leu Ser Leu Lys Asp Arg Glu Arg Arg Leu Gly
1 5 10 15
Phe Arg Gly Ala Val Ala Ala Asp Ala Arg Pro Leu Leu Ser Val Ala
20 25 30
Thr Ile Val Gly Tyr Pro Arg Xaa Val Ala Ile Gly Leu Cys Ser Tyr
35 40 45
Pro Arg Leu Pro Arg Lys Val Leu Cys Cys
50 55

(2) INFORMATION FOR SEQ ID NO:428:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 98 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide
(B) LOCATION: 1..98
(D) OTHER INFORMATION: / Ceres Seq. ID 1482240
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:428:
Arg Arg Pro Ser Ser Val Val Glu Gly Gln Gly Ala Ala Ala Arg Val
1 5 10 15
Ser Arg Cys Cys Gly Gly Arg Arg Pro Ser Ser Thr Ile Arg Gly Asp
20 25 30
His Arg Arg Leu Ser Ala Xaa Cys Gly Asp Arg Pro Val Leu Leu Ser
35 40 45
Ala Val Ala Glu Glu Ser Thr Met Leu Leu Met Leu Asp Ile Gln Ile
50 55 60
Gly Met Ser Thr Trp Pro Arg Thr Lys Asp Asn Cys Ile Met Phe Arg
65 70 75 80
Asn Gly Glu Met Ile Leu His Leu Met Ala His Ser Lys Val Lys His
85 90 95
Gly Lys

(2) INFORMATION FOR SEQ ID NO:429:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 798 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: DNA (genomic)
(ix) FEATURE:

(A) NAME/KEY: -
(B) LOCATION: 1..798
(D) OTHER INFORMATION: / Ceres Seq. ID 1482245
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:429:
aaccatagaa atcctccaat tattcgattc ccaagccact aaggcccttg gaggaacaac 60
agaatccaga aaccgaaaag argcctcact gctgccgcct gggccaagtc gtcgtcttgc 120
tttgccaatc cgtcgctcca cctgaacaca ccggcgaaga ggaggcgaag aagcgatggg 180
cgtgaccaag gaggacgtcg aggcggccat cacctctgct ctacagccctt ccaatctcgt 240
ggtgacggac acgtccggag ggtgtggcgc gagctacgag atcgaggtgg tgtcggagaa 300
gttcgagggg aagcggctgc tggagaggca ccggatggtg aacaccgcgc tggcgtctca 360
catggcggag atccacgccg tctccatcaa gaaggcgctc acccggctc aggccagacc 420
ccagggccca gccggagccg gccgccgata agccccaggc ttaagtgtt aacaccccc 480
aaaacggttt gatcccatat gccgatgcac gattacattg gctatctgct tgaataatgc 540
gggcggatgc acttgctaaa ttgcaggatg ttatccttga ctgattagaa acttctgcac 600
cgtgcattta acttctgtgt cactgtgtgt gtgttctgga tgcttctgcc ctggtcgttt 660
gctcgagact gtgtgttgca gttcatgctg ttaatgttct gccaggggtg ggttttcagt 720
cctggaattt ttatatttga ctgttgcctat gtctttcctt gcttgtaggg gtaaggggtt 780
tattctttaa ccttgttg

(2) INFORMATION FOR SEQ ID NO:430:
(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 47 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide
(ix) FEATURE:
(A) NAME/KEY: peptide
(B) LOCATION: 1..47
(D) OTHER INFORMATION: / Ceres Seq. ID 1482246
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:430:

Thr Ile Glu Ile Leu Gln Leu Phe Asp Ser Gln Ala Thr Lys Ala Leu
1 5 10 15
Gly Gly Thr Thr Glu Ser Arg Asn Arg Lys Xaa Ala Ser Leu Leu Pro

20 25 30
Pro Gly Pro Ser Arg Arg Leu Ala Leu Pro Ile Arg Arg Ser Thr
35 40 45

(2) INFORMATION FOR SEQ ID NO:431:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 91 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..91
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482247

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:431:

Met Gly Val Thr Lys Glu Asp Val Glu Ala Ala Ile Thr Ser Ala Leu
1 5 10 15
Ser Pro Ser Asn Leu Val Val Thr Asp Thr Ser Gly Gly Cys Gly Ala
20 25 30
Ser Tyr Glu Ile Glu Val Val Ser Glu Lys Phe Glu Gly Lys Arg Leu
35 40 45
Leu Glu Arg His Arg Met Val Asn Thr Ala Leu Ala Ser His Met Ala
50 55 60
Glu Ile His Ala Val Ser Ile Lys Lys Ala Leu Thr Pro Ala Gln Ala
65 70 75 80
Gln Pro Gln Gly Pro Ala Gly Ala Gly Arg Arg
85 90

(2) INFORMATION FOR SEQ ID NO:432:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 572 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..572
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482248

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:432:

atgggtggtc cgcactccgc accggtaccg cacaaccccc acaccgcag catccccatt 60
tctccgtccc aaaaccctag gctagtcccc ccacacctgg atccatcggg tcggaggcca 120
tgacgacggc gaggctccga tcctcggcct ccctccgcgg ggctctcctc cgccacttct 180
ccgtgggtcc cgctcgact ccgcgcggcg tctcccggtt ccagatttc caggttcctc 240
agtctattat gtggaggcat ttcgcaacgt ccaggcctaa ttctcttgca agacgcgaca 300
actttggtct gatggcctgt ttgcacgctc agatacgatg cgcttcgcag gctgctgctg 360
tgaaagaaac cgaatccagt agcagcaaga taagcatcgg gcccaccca aaacagatca 420
aggaggatga cgaggatgct aacctggtat accaagggcc aatatcatcg accataaaga 480
aagtgaagct tctctccctg tccacctgct gcctctccgt gtcgctgggg ccagtggtaa 540
cattcatgac ttgcctgac atgaatgtga tc

(2) INFORMATION FOR SEQ ID NO:433:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 151 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..151

(D) OTHER INFORMATION: / Ceres Seq. ID 1482249

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:433:

Met Thr Thr Ala Arg Leu Arg Ser Ser Ala Ser Leu Arg Gly Ala Leu
1 5 10 15
Leu Arg His Phe Ser Val Gly Pro Ala Ser Thr Pro Arg Ala Val Ser
20 25 30
Arg Val Pro Asp Phe Gln Val Pro Gln Ser Ile Met Trp Arg His Phe
35 40 45
Ala Thr Ser Arg Pro Asn Ser Leu Ala Arg Arg Asp Asn Phe Gly Leu
50 55 60
Met Ala Cys Leu His Ala Gln Ile Arg Cys Ala Ser Gln Ala Ala Ala
65 70 75 80
Val Lys Glu Thr Glu Ser Ser Ser Ser Lys Ile Ser Ile Gly Pro Lys
85 90 95
Pro Lys Gln Ile Lys Glu Asp Asp Glu Asp Ala Asn Leu Val Tyr Gln
100 105 110
Gly Pro Ile Ser Ser Thr Ile Lys Lys Val Lys Leu Leu Ser Leu Ser
115 120 125
Thr Cys Cys Leu Ser Val Ser Leu Gly Pro Val Val Thr Phe Met Thr
130 135 140
Ser Pro Asp Met Asn Val Ile
145 150

(2) INFORMATION FOR SEQ ID NO:434:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 108 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..108

(D) OTHER INFORMATION: / Ceres Seq. ID 1482250

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:434:

Met Trp Arg His Phe Ala Thr Ser Arg Pro Asn Ser Leu Ala Arg Arg
1 5 10 15
Asp Asn Phe Gly Leu Met Ala Cys Leu His Ala Gln Ile Arg Cys Ala
20 25 30
Ser Gln Ala Ala Ala Val Lys Glu Thr Glu Ser Ser Ser Lys Ile
35 40 45
Ser Ile Gly Pro Lys Pro Lys Gln Ile Lys Glu Asp Asp Glu Asp Ala
50 55 60
Asn Leu Val Tyr Gln Gly Pro Ile Ser Ser Thr Ile Lys Lys Val Lys
65 70 75 80
Leu Leu Ser Leu Ser Thr Cys Cys Leu Ser Val Ser Leu Gly Pro Val
85 90 95
Val Thr Phe Met Thr Ser Pro Asp Met Asn Val Ile
100 105

(2) INFORMATION FOR SEQ ID NO:435:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 87 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..87

(D) OTHER INFORMATION: / Ceres Seq. ID 1482251

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:435:

```
Met Ala Cys Leu His Ala Gln Ile Arg Cys Ala Ser Gln Ala Ala Ala
1           5           10           15
Val Lys Glu Thr Glu Ser Ser Ser Ser Lys Ile Ser Ile Gly Pro Lys
          20           25           30
Pro Lys Gln Ile Lys Glu Asp Asp Glu Asp Ala Asn Leu Val Tyr Gln
          35           40           45
Gly Pro Ile Ser Ser Thr Ile Lys Lys Val Lys Leu Leu Ser Leu Ser
          50           55           60
Thr Cys Cys Leu Ser Val Ser Leu Gly Pro Val Val Thr Phe Met Thr
65           70           75           80
Ser Pro Asp Met Asn Val Ile
          85
```

(2) INFORMATION FOR SEQ ID NO:436:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 519 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..519
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482254

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:436:

```
aaggctcact gctcagtgt cactgctcac tagctaaaaa catctccttt cttatccatg      60
gaganggcag cgctcacctc ccactccctg cagcgcccag cagcagcagc agctgctccc      120
gcgcatggcc agcngcggag ggtcggagcg gcgggcctgc ggcaccggca gccgcgcgcc      180
ggcggcagga tccgggccct gccttcggcg gaggtcatca gcgagatcct gagccccaag      240
ctggtgcccc gctcgcccg cgcaccggc gacgtctcct cgctcgctcc ggtcagtgcc      300
ctgatgtctg tcttctactt cgtgtccaac tgggtggtgc ccgagctgct cctgaagggc      360
ctcaacgagc ccaagcccga ggacgaagcg tccacgtcct tcgccgcgtc cgcgnacaac      420
gccgcccgcg ctggcccagc agacgacggc ggcaccgcta agatccgcct caaggtcaag      480
aagaagaaga acgggaaagc gaccatcgtc aaggtctag
```

(2) INFORMATION FOR SEQ ID NO:437:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 152 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..152
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482255

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:437:

```
Gly Ser Leu Leu Ser Ala His Cys Ser Leu Ala Lys Asn Ile Ser Phe
1           5           10           15
Leu Ile His Gly Xaa Gly Ser Ala His Leu Pro Leu Pro Ala Ala Pro
          20           25           30
Ser Ser Ser Ser Ser Cys Ser Arg Ala Trp Pro Xaa Ala Glu Gly Arg
          35           40           45
Ser Gly Gly Pro Ala Ala Pro Ala Ala Ala Arg Arg Arg Gln Asp Pro
          50           55           60
Gly Pro Ala Phe Gly Gly Gly His Gln Arg Asp Pro Glu Pro Gln Ala
65           70           75           80
Gly Ala Arg Leu Ala Arg Arg His Arg Arg Leu Leu Ala Arg Pro
          85           90           95
Gly Gln Cys Pro Asp Ala Ala Leu Leu Leu Arg Val Gln Leu Gly Gly
```

```

      100      105      110
Ala Arg Ala Ala Pro Glu Gly Pro Gln Arg Ala Gln Ala Arg Gly Arg
      115      120      125
Ser Val His Val Leu Arg Arg Val Arg Xaa Gln Arg Arg Arg Arg Trp
      130      135      140
Pro Ser Arg Arg Arg Arg His Arg
145      150
```

(2) INFORMATION FOR SEQ ID NO:438:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 153 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..153
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482256

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:438:

```

Met Glu Xaa Ala Ala Leu Thr Ser His Ser Leu Gln Arg Pro Ala Ala
1      5      10      15
Ala Ala Ala Ala Pro Ala His Gly Gln Xaa Arg Arg Val Gly Ala Ala
      20      25      30
Gly Leu Arg His Arg Gln Pro Arg Ala Gly Gly Arg Ile Arg Ala Leu
      35      40      45
Pro Ser Ala Glu Val Ile Ser Glu Ile Leu Ser Pro Lys Leu Val Pro
      50      55      60
Gly Ser Pro Ala Asp Thr Gly Asp Val Ser Ser Leu Val Pro Val Ser
      65      70      75      80
Ala Leu Met Leu Leu Phe Tyr Phe Val Ser Asn Trp Val Val Pro Glu
      85      90      95
Leu Leu Leu Lys Gly Leu Asn Glu Pro Lys Pro Glu Asp Glu Ala Ser
      100     105     110
Thr Ser Phe Ala Ala Ser Ala Xaa Asn Ala Ala Ala Gly Pro Ala
      115     120     125
Asp Asp Gly Gly Thr Gly Lys Ile Arg Leu Lys Val Lys Lys Lys Lys
      130     135     140
Asn Gly Lys Ala Thr Ile Val Lys Val
145     150
```

(2) INFORMATION FOR SEQ ID NO:439:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 278 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..278
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482257

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:439:

```

artgcaagca tatrgngcgc cgtgccagcc tgctcctcgc cgcrgcgctg ctcgtcgccg      60
tcgctgccgc ggcggtgccs cgacgtgcga gcgcacgcag tgcccggcgt acgaggtggt      120
ggacagcgcc aacgggttcg agatccggcg gtacacggac gccatgtgga tcaccacggc      180
gcccacgcag gacatctcct tcgtgcgcgc cgcgcgcacc ggcttcctac agctgttcga      240
ctacatchbag ggcaagaacg cgtacaacca gacgatcg
```

(2) INFORMATION FOR SEQ ID NO:440:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 92 amino acids

- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..92
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1482258

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:440:

Xaa	Ala	Ser	Ile	Xaa	Xaa	Ala	Val	Pro	Ala	Cys	Ser	Ser	Pro	Xaa	Arg
1			5						10					15	
Cys	Ser	Ser	Pro	Ser	Leu	Pro	Arg	Arg	Cys	Xaa	Asp	Val	Arg	Ala	His
			20					25					30		
Arg	Val	Pro	Gly	Val	Arg	Gly	Gly	Gly	Gln	Arg	Gln	Arg	Val	Arg	Asp
		35				40					45				
Pro	Ala	Val	His	Gly	Arg	His	Val	Asp	His	His	Gly	Ala	His	Arg	Gly
		50				55					60				
His	Leu	Leu	Arg	Arg	Arg	His	Ala	His	Arg	Leu	Pro	Thr	Ala	Val	Arg
65					70					75					80
Leu	His	Xaa	Gly	Gln	Glu	Arg	Val	Gln	Pro	Asp	Asp				
			85					90							

(2) INFORMATION FOR SEQ ID NO:441:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 92 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..92
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1482259

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:441:

Xaa	Gln	Ala	Tyr	Xaa	Ala	Pro	Cys	Gln	Pro	Ala	Pro	Arg	Arg	Xaa	Ala
1			5					10						15	
Ala	Arg	Arg	Arg	Arg	Cys	Arg	Gly	Gly	Ala	Xaa	Thr	Cys	Glu	Arg	Ile
			20					25					30		
Glu	Cys	Pro	Ala	Tyr	Glu	Val	Val	Asp	Ser	Ala	Asn	Gly	Phe	Glu	Ile
		35				40					45				
Arg	Arg	Tyr	Thr	Asp	Ala	Met	Trp	Ile	Thr	Thr	Ala	Pro	Ile	Glu	Asp
		50				55					60				
Ile	Ser	Phe	Val	Ala	Ala	Thr	Arg	Thr	Gly	Phe	Leu	Gln	Leu	Phe	Asp
65					70					75					80
Tyr	Ile	Xaa	Gly	Lys	Asn	Ala	Tyr	Asn	Gln	Thr	Ile				
			85					90							

(2) INFORMATION FOR SEQ ID NO:442:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 92 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..92
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1482260

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:442:

Cys	Lys	His	Xaa	Xaa	Arg	Arg	Ala	Ser	Leu	Leu	Leu	Ala	Xaa	Ala	Leu
1				5					10					15	

Leu Val Ala Val Ala Ala Ala Val Xaa Arg Arg Ala Ser Ala Ser
20 25 30
Ser Ala Arg Arg Thr Arg Trp Trp Thr Ala Pro Thr Gly Ser Arg Ser
35 40 45
Gly Gly Thr Arg Thr Pro Cys Gly Ser Pro Arg Arg Pro Ser Arg Thr
50 55 60
Ser Pro Ser Ser Pro Pro Arg Ala Pro Ala Ser Tyr Ser Cys Ser Thr
65 70 75 80
Thr Xaa Arg Ala Arg Thr Arg Thr Thr Arg Arg Ser
85 90

(2) INFORMATION FOR SEQ ID NO:443:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 931 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..931
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482261

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:443:

gaattcctcg	ccgccgtctt	cgtccaccag	aaccatggcc	tccgacaccg	cctcggcagt	60
tccgtcgtct	gtggtctcag	ctgccgagga	gacgctcgga	tataccgaat	ccgtagggac	120
catctctccc	atctgctcgc	ggcggcgggc	gcggaccccg	acgccgtggc	cgagctccca	180
ccccctctcc	gggcgcgcgc	tttccttgcc	ttggcgcagg	ccgcgacctc	ccttctcggg	240
gttcgtttta	gggtgttcggg	agttgaccct	gacgagcacc	ccatcagaaa	ggagtttgaa	300
aggttaagcc	taatgcagga	gaagttaa	caatttgaga	actgggacaa	agcaccactt	360
cgccttctta	ctacactaaa	tacacaagca	gcagcaaggt	tcattggaca	ctcactttcc	420
catctgacat	ctgatcagaa	gaggagcatg	catgaaataa	gtagaggaga	aaggcggagt	480
tgggtctgggc	agaagagaaa	gcctgaacct	tcagtagaaa	agaagtctgt	tcgtgctgct	540
gcagaagagt	tccttgcaaa	ggcttctcag	gaacttattg	gacatagtga	tagcagggtc	600
aagggtcctg	ttatactcat	ttctcgatga	gatgaggact	agatcaaaaa	aatgggcgct	660
taccagatta	catgcctgat	tcctcggtga	ggcaaaggaa	ggtagaagtt	cctggtgatg	720
aagataaact	tacgtacatt	gctgtggtga	tgaagatgaa	tttatctgca	ttgctgtgtt	780
gttctacatg	taacagggaa	tggagcaaag	ctgcataggc	ttgcttaagt	ccccagttct	840
gggagcaatt	ggcctcgaat	cttgagtgc	atttatctga	gtttctttcc	ggaaagaatt	900
ttgacattct	atttgctagt	ggaactggag	c			

(2) INFORMATION FOR SEQ ID NO:444:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 213 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..213
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482262

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:444:

Glu Phe Leu Ala Ala Val Phe Val His Gln Asn His Gly Leu Arg His
1 5 10 15
Arg Leu Gly Ser Ser Val Val Cys Gly Leu Ser Cys Arg Gly Asp Ala
20 25 30
Arg Ile Tyr Arg Ile Arg Arg Asp His Leu Ser His Leu Leu Ala Ala
35 40 45
Ala Ala Ala Asp Pro Asp Ala Val Ala Glu Leu Pro Pro Leu Leu Arg
50 55 60
Ala Arg Ala Phe Leu Ala Leu Ala Gln Ala Ala Thr Ser Leu Leu Gly

(2) INFORMATION FOR SEQ ID NO:445:

(A) LENGTH: 109 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME

(B) LOCATION: 1..109

(D) OTHER INFORMATION

SEQUENCE DESCRIPTION: SEQ ID NO:445:

Gln Glu Lys Leu Asn Gln Phe Glu Asn Trp A

(2) INFORMATION FOR SEQ ID NO:446:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 600 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..600

(D) OTHER INFORMATION: / Ceres Seq. ID 1482264

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:446:

60
120

cctgctcgtc	tcctcatggg	attcggggct	gcggttgtag	gatgccgacg	agggcacgct	180
cagggtcaac	gtggagtcag	aggcggcatt	cctcgactgc	tgcttcgagg	atgagttctgc	240
agcgtttgcc	tgcggtctctg	atggatctgt	gagaaggtag	gacttccact	cagggttcgca	300
ggatacgggtg	gggctccatg	aagatgcact	agcctgcatt	gagttctctt	cactgaccgg	360
tcagattatg	acaggcagcc	ttgacaagaa	gctaaagctt	tgggattcaa	aaacaagaaa	420
tgtaagcccg	agcggcacca	taaccttaaa	ttcagatgtg	gcctcaattt	ctatatgcgg	480
catttacata	ttagctgcag	ttgagagaaa	tgtttatctt	tatgacatga	ggaatctaac	540
aagaccagtt	gatgaaaaaa	gactgtcctc	tggattatca	aattcgatgc	cttcatactt	600

(2) INFORMATION FOR SEQ ID NO:447:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 199 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..199
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482265

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:447:

Leu	Thr	Ala	Arg	Arg	Xaa	Ala	Ala	Pro	Arg	Ser	Xaa	Pro	Val	Asp	Gly	
1			5					10						15		
Glu	Ala	Glu	Leu	Ala	Val	Asp	Ala	Thr	Ala	Gly	Ala	Ala	Ser	Arg	Val	
			20					25					30			
Arg	Phe	Ala	Pro	Thr	Ser	Asn	Asn	Leu	Leu	Val	Ser	Ser	Trp	Asp	Ser	
			35				40					45				
Gly	Leu	Arg	Leu	Tyr	Asp	Ala	Asp	Glu	Gly	Thr	Leu	Arg	Val	Asn	Val	
			50			55				60						
Glu	Ser	Glu	Ala	Ala	Phe	Leu	Asp	Cys	Cys	Phe	Glu	Asp	Glu	Ser	Ala	
65					70					75					80	
Ala	Phe	Ala	Cys	Gly	Ser	Asp	Gly	Ser	Val	Arg	Arg	Tyr	Asp	Phe	His	
			85						90					95		
Ser	Gly	Ser	Gln	Asp	Thr	Val	Gly	Leu	His	Glu	Asp	Ala	Leu	Ala	Cys	
			100					105					110			
Ile	Glu	Phe	Ser	Ser	Leu	Thr	Gly	Gln	Ile	Met	Thr	Gly	Ser	Leu	Asp	
			115				120					125				
Lys	Lys	Leu	Lys	Leu	Trp	Asp	Ser	Lys	Thr	Arg	Asn	Val	Ser	Pro	Ser	
			130			135					140					
Gly	Thr	Ile	Thr	Leu	Asn	Ser	Asp	Val	Ala	Ser	Ile	Ser	Ile	Cys	Gly	
145					150					155					160	
Ile	Tyr	Ile	Leu	Ala	Ala	Val	Glu	Arg	Asn	Val	Tyr	Leu	Tyr	Asp	Met	
			165						170					175		
Arg	Asn	Leu	Thr	Arg	Pro	Val	Asp	Glu	Lys	Arg	Leu	Ser	Ser	Gly	Leu	
			180					185					190			
Ser	Asn	Ser	Met	Pro	Ser	Tyr										
			195													

(2) INFORMATION FOR SEQ ID NO:448:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 516 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..516
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482270

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:448:

```
gcgtcaccat ccatttgacg aggcgggttta tccccagca ccccaaccaa cctttccacg      60
taccaccggg tttctgtccg cgccccgccc ttcaaaagca ggtccgcacg ccggccggcg      120
agacagacga caccaccacg ccgggacggg aggcacaggt gcggtctgcg tcgagagttg      180
gtccactggc aggcgggaat gaagaagtgc gcgtcggagc tggagctgga ggcgttcacg      240
cgggagagcg gcgaggacgc ccgcgccgcc gccggaggta gcagtccggg gtgcggtgga      300
tcaagcgatc ccggagggag cggcgtcttc tcaccgggct tcggtttcgc cgactcggac      360
accatggatg gaggcagttg gtggtacggg aacgtccgca cgccgaaccc agtcatgtcg      420
caggcggcgt ccatatccgc tagccccggg ctaaccacct cagccaatca tgctcttgaa      480
agcgagtcag actccgacag cgaatcactg tatgag
```

(2) INFORMATION FOR SEQ ID NO:449:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 66 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..66
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482271

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:449:

```
Arg His His Pro Phe Asp Glu Ala Val Tyr Pro Pro Ala Pro Gln Pro
1          5          10          15
Thr Phe Pro Arg Thr Thr Gly Phe Leu Ser Ala Pro Arg Pro Ser Lys
          20          25          30
Ala Gly Pro His Ala Gly Arg Arg Asp Arg Arg His His His Ala Gly
          35          40          45
Thr Gly Gly Thr Gly Ala Val Cys Val Glu Ser Trp Ser Thr Gly Arg
50          55          60
Pro Glu
65
```

(2) INFORMATION FOR SEQ ID NO:450:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 92 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..92
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482272

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:450:

```
Val Thr Ile His Leu Thr Arg Arg Phe Ile Pro Gln His Pro Asn Gln
1          5          10          15
Pro Phe His Val Pro Pro Gly Phe Cys Pro Arg Pro Ala Leu Gln Lys
          20          25          30
Gln Val Arg Thr Pro Ala Gly Glu Thr Asp Asp Thr Thr Thr Pro Gly
          35          40          45
Arg Glu Ala Gln Val Arg Ser Ala Ser Arg Val Gly Pro Leu Ala Gly
50          55          60
Arg Asn Glu Glu Val Arg Val Gly Ala Gly Ala Gly Gly Val His Pro
65          70          75          80
Gly Glu Arg Arg Gly Arg Pro Arg Arg Arg Arg Arg
          85          90
```

(2) INFORMATION FOR SEQ ID NO:451:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 106 amino acids
- (B) TYPE: amino acid

- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..106
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1482273
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:451:

```
Met Lys Lys Cys Ala Ser Glu Leu Glu Leu Glu Ala Phe Ile Arg Glu
 1             5             10             15
Ser Gly Glu Asp Ala Arg Ala Ala Gly Gly Ser Ser Pro Gly Cys
          20             25             30
Gly Gly Ser Ser Asp Pro Gly Gly Ser Gly Val Phe Ser Pro Gly Phe
          35             40             45
Gly Phe Ala Asp Ser Asp Thr Met Asp Gly Gly Ser Trp Trp Tyr Gly
          50             55             60
Asn Val Arg Thr Pro Asn Pro Val Met Ser Gln Ala Ala Ser Ile Ser
65             70             75             80
Ala Ser Pro Gly Leu Thr Thr Ser Ala Asn His Ala Leu Glu Ser Glu
          85             90             95
Ser Asp Ser Asp Ser Glu Ser Leu Tyr Glu
          100             105
```

- (2) INFORMATION FOR SEQ ID NO:452:
 - (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 561 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear
 - (ii) MOLECULE TYPE: DNA (genomic)
 - (ix) FEATURE:
 - (A) NAME/KEY: -
 - (B) LOCATION: 1..561
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1482274
 - (xi) SEQUENCE DESCRIPTION: SEQ ID NO:452:

```
aagatggaca ggctctgtcg ccactgctac accagtacac cctgcacgcg ctgcggggcgt      60
ttccgccggc tttgtgcctc ctctgccctc cccggggcgt cgccctccgt ccacgctcaa      120
gctcgctccg tcccggcgcc tcgaactcgt cgtcctcgct tccgctgtcg ccaccgcgaa      180
gcatgaggag gcgtcctggg atcactggct tgcagaatgt ggcggtact ctcgaactat      240
cagaaccagt tgggactggt cggggacaat atggccaagg tccggaccga tgtcatgaag      300
aagcagcgac ttgggatggt ccgatcacag ctcgagaaat ttgcttgcaa gcataaggtt      360
ttgagcaggt ttggtgcaat ctgattttga acctgctatg gacatcttcc actcaagttc      420
ttgtcaatgg agtgcctgga cagcctataa atctcaagcg tgggttgcaa caagctgctc      480
tgaaacatcc attttctctg kgccatcaat ggatgcttag cacagatcgc ttccacatat      540
cacatgatca atcagtactg g
```

- (2) INFORMATION FOR SEQ ID NO:453:
 - (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 61 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear
 - (ii) MOLECULE TYPE: peptide
 - (ix) FEATURE:
 - (A) NAME/KEY: peptide
 - (B) LOCATION: 1..61
 - (D) OTHER INFORMATION: / Ceres Seq. ID 1482275
 - (xi) SEQUENCE DESCRIPTION: SEQ ID NO:453:

```
Lys Met Asp Arg Leu Cys Arg His Cys Tyr Thr Ser Thr Pro Cys Thr
 1             5             10             15
Arg Cys Gly Arg Phe Arg Arg Leu Cys Ala Ser Ser Ala Leu Pro Gly
```

	20		25		30										
Arg	Ser	Pro	Ser	Val	His	Ala	Gln	Ala	Arg	Ser	Val	Pro	Ala	Pro	Arg
	35						40						45		
Thr	Arg	Arg	Pro	Arg	Phe	Arg	Cys	Arg	His	Arg	Glu	Ala			
	50					55					60				

(2) INFORMATION FOR SEQ ID NO:454:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 98 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..98

(D) OTHER INFORMATION: / Ceres Seq. ID 1482276

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:454:

Arg	Trp	Thr	Gly	Ser	Val	Ala	Thr	Ala	Thr	Pro	Val	His	Pro	Ala	Arg
1				5					10					15	
Ala	Ala	Gly	Val	Ser	Ala	Gly	Phe	Val	Pro	Pro	Leu	Pro	Ser	Pro	Gly
			20					25					30		
Ala	Arg	Pro	Pro	Ser	Thr	Leu	Lys	Leu	Ala	Pro	Ser	Arg	Arg	Leu	Glu
	35					40						45			
Leu	Val	Val	Leu	Ala	Ser	Ala	Val	Ala	Thr	Ala	Lys	His	Glu	Glu	Ala
	50					55					60				
Ser	Trp	Asp	His	Trp	Leu	Ala	Glu	Cys	Gly	Gly	Tyr	Ser	Arg	Thr	Ile
65					70					75					80
Arg	Thr	Ser	Trp	Asp	Trp	Ser	Gly	Thr	Ile	Trp	Pro	Arg	Ser	Gly	Pro
				85					90					95	

Met Ser

(2) INFORMATION FOR SEQ ID NO:455:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 117 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..117

(D) OTHER INFORMATION: / Ceres Seq. ID 1482277

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:455:

Asp	Gly	Gln	Ala	Leu	Ser	Pro	Leu	Leu	His	Gln	Tyr	Thr	Leu	His	Ala
1				5					10					15	
Leu	Arg	Ala	Phe	Pro	Pro	Ala	Leu	Cys	Leu	Leu	Cys	Pro	Pro	Arg	Ala
			20					25					30		
Leu	Ala	Leu	Arg	Pro	Arg	Ser	Ser	Leu	Arg	Pro	Gly	Ala	Ser	Asn	
	35					40					45				
Ser	Ser	Ser	Ser	Leu	Pro	Leu	Ser	Pro	Pro	Arg	Ser	Met	Arg	Arg	Arg
	50					55				60					
Pro	Gly	Ile	Thr	Gly	Leu	Gln	Asn	Val	Ala	Ala	Thr	Leu	Glu	Leu	Ser
65					70					75				80	
Glu	Pro	Val	Gly	Thr	Gly	Arg	Gly	Gln	Tyr	Gly	Gln	Gly	Arg	Asp	Arg
				85				90					95		
Cys	His	Glu	Glu	Ala	Ala	Thr	Trp	Asp	Gly	Pro	Ile	Thr	Ala	Arg	Glu
			100					105					110		

Ile Cys Leu Gln Ala
115

(2) INFORMATION FOR SEQ ID NO:456:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 578 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..578
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482282

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:456:

```
acctcggctc gggcgagag cgcgcggcgc cggcgcgctc tctcctcctg ctccgatctc 60
tctgccagc cccgcctgtg cgcttcgcta ctacgcttcc tccatgcac ccctcagcca 120
tgtaagcctg tatttgggaa gacattagcg gagttgaatc cagaagaaga gccgaagagt 180
tatcttagcc acagccaggt cgcccgttag ttgttcgcgg aaatgtccct ccgccagctg 240
cttcacaaaa cgcgtccgtg gcgcgcgctt gaggagccca cgaagatgtc ttgtctcctc 300
tccatcttcc gtgcgtcttc cattctcgtt tctgaaggct cggctgagcc actgcgccga 360
tcttcatctg tgccagcccc gctgccaaaga agcttgctt gctccagctc tgacccccctt 420
ggccccagat tcagcatcga cgtggtcgac tcggaccatt ggccctcgtc atttgacttg 480
ktgtccgacg ctgcacggag caatgaatgc ccagatgtct ncgagcaaca tgaggatgat 540
gaactgcmcg actcttatga tgagatagat gacatgag
```

(2) INFORMATION FOR SEQ ID NO:457:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 69 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..69
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482283

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:457:

```
Thr Ser Ala Arg Ala Gln Ser Ala Arg Arg Arg Arg Val Leu Ser Ser
1           5           10           15
Cys Ser Asp Leu Ser Ala Gln Pro Arg Leu Cys Ala Ser Leu Leu Arg
          20           25           30
Phe Leu His Ala Ser Pro Gln Pro Cys Thr Pro Val Phe Gly Lys Thr
          35           40           45
Leu Ala Glu Leu Asn Pro Glu Glu Glu Pro Lys Ser Tyr Leu Ser His
          50           55           60
Ser Gln Val Ala Arg
65
```

(2) INFORMATION FOR SEQ ID NO:458:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 118 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..118
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482284

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:458:

```
Met Ser Leu Arg Gln Leu Leu His Gln Thr Arg Pro Trp Arg Ala Leu
1           5           10           15
Glu Gln Pro Thr Lys Met Ser Cys Leu Leu Ser Ile Phe Arg Ala Leu
```


20 25 30
Ser Ile Leu Arg Ser Glu Gly Ser Ala Glu Pro Leu Arg Arg Ser Ser
35 40 45
Ser Val Pro Ala Pro Leu Pro Arg Ser Leu Pro Cys Ser Ser Ser Asp
50 55 60
Pro Leu Gly Pro Arg Phe Ser Ile Asp Val Val Asp Ser Asp His Trp
65 70 75 80
Pro Ser Ser Phe Asp Leu Xaa Ser Asp Ala Ala Arg Ser Asn Glu Cys
85 90 95
Pro Asp Val Xaa Glu Gln His Glu Asp Asp Glu Leu Xaa Asp Ser Tyr
100 105 110
Asp Glu Ile Asp Asp Met
115

(2) INFORMATION FOR SEQ ID NO:459:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 97 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..97
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482285

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:459:

Met Ser Cys Leu Leu Ser Ile Phe Arg Ala Leu Ser Ile Leu Arg Ser
1 5 10 15
Glu Gly Ser Ala Glu Pro Leu Arg Arg Ser Ser Ser Val Pro Ala Pro
20 25 30
Leu Pro Arg Ser Leu Pro Cys Ser Ser Ser Asp Pro Leu Gly Pro Arg
35 40 45
Phe Ser Ile Asp Val Val Asp Ser Asp His Trp Pro Ser Ser Phe Asp
50 55 60
Leu Xaa Ser Asp Ala Ala Arg Ser Asn Glu Cys Pro Asp Val Xaa Glu
65 70 75 80
Gln His Glu Asp Asp Glu Leu Xaa Asp Ser Tyr Asp Glu Ile Asp Asp
85 90 95
Met

(2) INFORMATION FOR SEQ ID NO:460:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 881 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..881
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482289

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:460:

tctctctttt ttccccagca atgcaattcc gcagacagac gcaggcgcca rgcggcaggc 60
ggcggcgcac cgcaccgctt cttcctcttc tatctctcat ctacagcctt cgctgcgccg 120
ccatggccac caccgcttg ctgccgtgc tccgacgccg cctcgccgcc gcaatgcgcg 180
gatcgctgc tccctactcc ctccgaggac cctcatttcc tgcaccagca gctgcagggc 240
taaggctccct cctaacagtt gctggagcga gcgatactgc aacagaaccc caggaccaac 300
agcattccga aacaactccc ccgccggctt ctgtcccgac accggagtcc ggtctcaaag 360
tcagggacac ctccaacctg aagatctcac caaggcatga cctcgccatg atctttacgt 420
gcaaggtgtg cgagaccagg tccatgaaga tggccagcag ggactcgtac gagaacggag 480

```
tcgtggctcgt gcggtgcggt ggctgcaaca acctccacct catggcggac aggcttggct 540
ggtttgggga gccagggagc atcgaggact tcctagcgac gcaaggggag gaggtgaaga 600
aagggttcgac agatactatc agctttactt tggacgactt ggctgggtct caggtcagtt 660
ctaagggggcc ttccgaacaa aattaatatg atagtgtttg gtccagtaag aacctgtaga 720
agcctctctt tactataaag aagatgcgcg tgtcacctgt gtgttgaaga aaaaaacgcc 780
tctagaagcc taccttaact gttgcacctg tagttctgct taacttcatg gcttttcatg 840
tgtagctttc gagcccatca aatatgcat gttgttatc t
```

(2) INFORMATION FOR SEQ ID NO:461:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 227 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..227

(D) OTHER INFORMATION: / Ceres Seq. ID 1482290

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:461:

```
Ser Leu Ser Ser Pro Ala Met Gln Phe Arg Arg Gln Thr Gln Ala Ala
1          5          10          15
Xaa Gly Arg Arg Arg Arg Thr Ala Pro Leu Leu Pro Leu Leu Ser Leu
20          25          30
Ile Tyr Ser Leu Arg Cys Ala Ala Met Ala Thr Thr Arg Leu Leu Pro
35          40          45
Leu Leu Arg Arg Arg Leu Ala Ala Ala Ile Ala Gly Ser Pro Ala Pro
50          55          60
Tyr Ser Leu Arg Gly Pro Ser Phe Pro Ala Pro Ala Ala Ala Gly Leu
65          70          75          80
Arg Ser Leu Leu Thr Val Ala Gly Ala Ser Asp Thr Ala Thr Glu Pro
85          90          95
Gln Asp Gln Gln His Ser Glu Thr Thr Pro Pro Pro Ala Ser Val Pro
100         105         110
Thr Pro Glu Ser Gly Leu Lys Val Arg Asp Thr Ser Asn Leu Lys Ile
115         120         125
Ser Pro Arg His Asp Leu Ala Met Ile Phe Thr Cys Lys Val Cys Glu
130         135         140
Thr Arg Ser Met Lys Met Ala Ser Arg Asp Ser Tyr Glu Asn Gly Val
145         150         155         160
Val Val Val Arg Cys Gly Gly Cys Asn Asn Leu His Leu Met Ala Asp
165         170         175
Arg Leu Gly Trp Phe Gly Glu Pro Gly Ser Ile Glu Asp Phe Leu Ala
180         185         190
Thr Gln Gly Glu Glu Val Lys Lys Gly Ser Thr Asp Thr Ile Ser Phe
195         200         205
Thr Leu Asp Asp Leu Ala Gly Ser Gln Val Ser Ser Lys Gly Pro Ser
210         215         220
Glu Gln Asn
225
```

(2) INFORMATION FOR SEQ ID NO:462:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 221 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..221

(D) OTHER INFORMATION: / Ceres Seq. ID 1482291

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:462:

Met	Gln	Phe	Arg	Arg	Gln	Thr	Gln	Ala	Ala	Xaa	Gly	Arg	Arg	Arg	Arg	
1				5					10					15		
Thr	Ala	Pro	Leu	Leu	Pro	Leu	Leu	Ser	Leu	Ile	Tyr	Ser	Leu	Arg	Cys	
			20					25					30			
Ala	Ala	Met	Ala	Thr	Thr	Arg	Leu	Leu	Pro	Leu	Leu	Arg	Arg	Arg	Leu	
			35				40					45				
Ala	Ala	Ala	Ile	Ala	Gly	Ser	Pro	Ala	Pro	Tyr	Ser	Leu	Arg	Gly	Pro	
			50			55				60						
Ser	Phe	Pro	Ala	Pro	Ala	Ala	Ala	Gly	Leu	Arg	Ser	Leu	Leu	Thr	Val	
65					70				75					80		
Ala	Gly	Ala	Ser	Asp	Thr	Ala	Thr	Glu	Pro	Gln	Asp	Gln	Gln	His	Ser	
				85				90						95		
Glu	Thr	Thr	Pro	Pro	Pro	Ala	Ser	Val	Pro	Thr	Pro	Glu	Ser	Gly	Leu	
			100					105					110			
Lys	Val	Arg	Asp	Thr	Ser	Asn	Leu	Lys	Ile	Ser	Pro	Arg	His	Asp	Leu	
			115				120					125				
Ala	Met	Ile	Phe	Thr	Cys	Lys	Val	Cys	Glu	Thr	Arg	Ser	Met	Lys	Met	
			130			135					140					
Ala	Ser	Arg	Asp	Ser	Tyr	Glu	Asn	Gly	Val	Val	Val	Val	Arg	Cys	Gly	
145					150				155					160		
Gly	Cys	Asn	Asn	Leu	His	Leu	Met	Ala	Asp	Arg	Leu	Gly	Trp	Phe	Gly	
				165				170						175		
Glu	Pro	Gly	Ser	Ile	Glu	Asp	Phe	Leu	Ala	Thr	Gln	Gly	Glu	Glu	Val	
			180				185						190			
Lys	Lys	Gly	Ser	Thr	Asp	Thr	Ile	Ser	Phe	Thr	Leu	Asp	Asp	Leu	Ala	
			195				200				205					
Gly	Ser	Gln	Val	Ser	Ser	Lys	Gly	Pro	Ser	Glu	Gln	Asn				
			210			215					220					

(2) INFORMATION FOR SEQ ID NO:463:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 187 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..187

(D) OTHER INFORMATION: / Ceres Seq. ID 1482292

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:463:

Met	Ala	Thr	Thr	Arg	Leu	Leu	Pro	Leu	Leu	Arg	Arg	Arg	Leu	Ala	Ala	
1				5					10					15		
Ala	Ile	Ala	Gly	Ser	Pro	Ala	Pro	Tyr	Ser	Leu	Arg	Gly	Pro	Ser	Phe	
			20					25					30			
Pro	Ala	Pro	Ala	Ala	Ala	Gly	Leu	Arg	Ser	Leu	Leu	Thr	Val	Ala	Gly	
			35				40					45				
Ala	Ser	Asp	Thr	Ala	Thr	Glu	Pro	Gln	Asp	Gln	Gln	His	Ser	Glu	Thr	
			50			55				60						
Thr	Pro	Pro	Pro	Ala	Ser	Val	Pro	Thr	Pro	Glu	Ser	Gly	Leu	Lys	Val	
65					70				75					80		
Arg	Asp	Thr	Ser	Asn	Leu	Lys	Ile	Ser	Pro	Arg	His	Asp	Leu	Ala	Met	
				85				90						95		
Ile	Phe	Thr	Cys	Lys	Val	Cys	Glu	Thr	Arg	Ser	Met	Lys	Met	Ala	Ser	
			100					105					110			
Arg	Asp	Ser	Tyr	Glu	Asn	Gly	Val	Val	Val	Arg	Cys	Gly	Gly	Cys		
			115				120					125				
Asn	Asn	Leu	His	Leu	Met	Ala	Asp	Arg	Leu	Gly	Trp	Phe	Gly	Glu	Pro	

130 135 140
Gly Ser Ile Glu Asp Phe Leu Ala Thr Gln Gly Glu Glu Val Lys Lys
145 150 155 160
Gly Ser Thr Asp Thr Ile Ser Phe Thr Leu Asp Asp Leu Ala Gly Ser
165 170 175
Gln Val Ser Ser Lys Gly Pro Ser Glu Gln Asn
180 185

(2) INFORMATION FOR SEQ ID NO:464:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 671 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..671
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482293

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:464:

gctcttttccc ccttgccccc ttcccagttc cactctgag cactctcctc cgctctgctc	60
ctttgtcccc caccgcaaac cgtaaaccct agcctgaggg gcacccctgt cgcagccatg	120
ggcgccascg gaagctgcag ggcgagatcg accgcgtcct gaagaaggtc caggaggcg	180
tcgatgtctt tgacagcatc tggaataagg tctacgacac tgagaatgcc aaccagaagg	240
agaagttcga ggcggacctc aagaaggaga tcaagaagct gcagcggnta cagggaccag	300
atcaagacgt ggattcagtc cagcgagatc aaggacaaga aggtctctgat ggatgctcga	360
aagcagattg aacgagagat ggaacgattt aaagtatgtg agaaggaaac aaaaactaag	420
gcattctcaa aagaagggtt aggtcagcaa ccaaaaacag atcccaaaga aaaggccaaa	480
gctgaaacaa gagactggct taataatgtg gtgtgttgga atcctgaatt gctactctta	540
tgctcttatg ttttcataatc tgttttttgg tatactaact gaaccacact gttaaactcg	600
aacatatgta tactattttg tttgagaata ccttggatct ttaattcatt tccgaggaca	660
tggtttgtgt c	

(2) INFORMATION FOR SEQ ID NO:465:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 132 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..132
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482294

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:465:

Met Ser Leu Thr Ala Ser Gly Ile Arg Ser Thr Thr Leu Arg Met Pro	
1 5 10 15	
Thr Arg Arg Arg Ser Ser Arg Arg Thr Ser Arg Arg Arg Ser Arg Ser	
20 25 30	
Cys Ser Xaa Tyr Arg Asp Gln Ile Lys Thr Trp Ile Gln Ser Ser Glu	
35 40 45	
Ile Lys Asp Lys Lys Ala Leu Met Asp Ala Arg Lys Gln Ile Glu Arg	
50 55 60	
Glu Met Glu Arg Phe Lys Val Cys Glu Lys Glu Thr Lys Thr Lys Ala	
65 70 75 80	
Phe Ser Lys Glu Gly Leu Gly Gln Gln Pro Lys Thr Asp Pro Lys Glu	
85 90 95	
Lys Ala Lys Ala Glu Thr Arg Asp Trp Leu Asn Asn Val Val Cys Trp	
100 105 110	
Asn Pro Glu Leu Leu Leu Cys Ser Tyr Val Phe Ile Ser Val Phe	
115 120 125	

Trp Tyr Thr Asn

130

(2) INFORMATION FOR SEQ ID NO:466:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 118 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..118

(D) OTHER INFORMATION: / Ceres Seq. ID 1482295

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:466:

Met Pro Thr Arg Arg Ser Ser Arg Arg Thr Ser Arg Arg Ser
1 5 10 15
Arg Ser Cys Ser Xaa Tyr Arg Asp Gln Ile Lys Thr Trp Ile Gln Ser
20 25 30
Ser Glu Ile Lys Asp Lys Lys Ala Leu Met Asp Ala Arg Lys Gln Ile
35 40 45
Glu Arg Glu Met Glu Arg Phe Lys Val Cys Glu Lys Glu Thr Lys Thr
50 55 60
Lys Ala Phe Ser Lys Glu Gly Leu Gly Gln Gln Pro Lys Thr Asp Pro
65 70 75 80
Lys Glu Lys Ala Lys Ala Glu Thr Arg Asp Trp Leu Asn Asn Val Val
85 90 95
Cys Trp Asn Pro Glu Leu Leu Leu Leu Cys Ser Tyr Val Phe Ile Ser
100 105 110
Val Phe Trp Tyr Thr Asn
115

(2) INFORMATION FOR SEQ ID NO:467:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 77 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..77

(D) OTHER INFORMATION: / Ceres Seq. ID 1482296

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:467:

Met Asp Ala Arg Lys Gln Ile Glu Arg Glu Met Glu Arg Phe Lys Val
1 5 10 15
Cys Glu Lys Glu Thr Lys Thr Lys Ala Phe Ser Lys Glu Gly Leu Gly
20 25 30
Gln Gln Pro Lys Thr Asp Pro Lys Glu Lys Ala Lys Ala Glu Thr Arg
35 40 45
Asp Trp Leu Asn Asn Val Val Cys Trp Asn Pro Glu Leu Leu Leu Leu
50 55 60
Cys Ser Tyr Val Phe Ile Ser Val Phe Trp Tyr Thr Asn
65 70 75

(2) INFORMATION FOR SEQ ID NO:468:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 868 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..868

(D) OTHER INFORMATION: / Ceres Seq. ID 1482297

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:468:

gtccatgcat	gggcatggaa	tggatggatg	tgaatgccac	gaacgattcc	gccccgccgg	60
ccaggtgaga	gatgagcctc	acggcggcct	tccgtgccac	caaaatcccg	cgcgctcttc	120
ctccaaagtg	cggtgagcct	gccgcctcct	cttcggcctc	ggcgtccggg	gatccgccgc	180
cggggggccg	gaagagtact	aaggcgccgc	cgcctgtgtg	cgtgtacctt	atagcctcat	240
cccggatccg	ccgcacgtac	gtcggcggtca	ccaccgattt	ccctcgccgg	ctgcggcaac	300
ataatggtga	gttaaaaggt	ggtgcaaaaag	cttcctctgc	cggcaggcct	tggaatctcg	360
catgccttgt	tgaaggattt	gccaacagaa	gtgaagcctg	tgagtttgaa	tcgaaatgga	420
agatcgtctc	ccgaaaaatt	gcacggaaaa	gaactgagct	tagcatgaag	tcagtgtctg	480
aacatcgaga	agcagctttg	agcagagtgg	aaacattcat	ggattgtagc	cacctaaaaa	540
tcaaatggca	gtcaagttga	gaccatttaa	tcacttgcac	tatgcagggtg	gcaggcatct	600
aacttgagga	aacatcacca	cttaagaatc	ctcctgtctt	ctagcagctc	gtagcaaaga	660
taacttataa	tcttctgctg	aaccatcaag	atggctgctg	ctatgctttc	ttaacatgaa	720
aaaccaagag	tagccccagt	ggaattctat	gtttgatttt	tttttctatg	aacaattggt	780
tccgaacaat	aatatggatc	atgcgacacc	cgtttgtaaa	tgtaaattat	acttatgtat	840
tgtaatcacc	tatatttctt	ctcattct				

(2) INFORMATION FOR SEQ ID NO:469:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 162 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..162

(D) OTHER INFORMATION: / Ceres Seq. ID 1482298

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:469:

Met	Ser	Leu	Thr	Ala	Phe	Arg	Ala	Thr	Lys	Ile	Pro	Arg	Ala	Leu
1			5					10					15	
Pro	Pro	Lys	Cys	Gly	Glu	Pro	Ala	Ala	Ser	Ser	Ser	Ala	Ser	Ala
			20					25					30	
Gly	Asp	Pro	Pro	Pro	Gly	Ala	Val	Lys	Ser	Thr	Lys	Ala	Pro	Pro
			35				40					45		
Trp	Cys	Val	Tyr	Leu	Ile	Ala	Ser	Ser	Arg	Ile	Arg	Arg	Thr	Tyr
	50					55					60			Val
Gly	Val	Thr	Thr	Asp	Phe	Pro	Arg	Arg	Leu	Arg	Gln	His	Asn	Gly
65					70					75				80
Leu	Lys	Gly	Gly	Ala	Lys	Ala	Ser	Ser	Ala	Gly	Arg	Pro	Trp	Asn
				85					90					95
Ala	Cys	Leu	Val	Glu	Gly	Phe	Ala	Asn	Arg	Ser	Glu	Ala	Cys	Glu
			100					105					110	Phe
Glu	Ser	Lys	Trp	Lys	Ile	Val	Ser	Arg	Lys	Ile	Ala	Arg	Lys	Arg
		115				120						125		Thr
Glu	Leu	Ser	Met	Lys	Ser	Val	Leu	Gln	His	Arg	Glu	Ala	Ala	Leu
	130					135					140			Ser
Arg	Val	Glu	Thr	Phe	Met	Asp	Cys	Ser	His	Leu	Lys	Ile	Lys	Trp
145					150					155				160
Ser	Ser													

(2) INFORMATION FOR SEQ ID NO:470:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 642 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

- (D) TOPOLOGY: linear
(ii) MOLECULE TYPE: DNA (genomic)
(ix) FEATURE:
 (A) NAME/KEY: -
 (B) LOCATION: 1..642
 (D) OTHER INFORMATION: / Ceres Seq. ID 1482299

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:470:

aaatttttct ccagccgccc cgctcctgat ccttatctct gcgcgcgctg catcggcgcc	60
cgccgggagg gagtcccgcc cgctcgtcc atgttggtggg tccgcaatat ccgccgcttc	120
gtcgacacgg gcgcggcct cggatccgag gccatcatgg aactggagac taaaaggata	180
ttgcttgaga ttttcaagga gcggcagcgn gaagagtgcc gaggctggtt ccatccaag	240
tttttacaag aaacctgaag aaggatccat tagctctaga gttcaaagggt tggccaagta	300
cagggtttcta aagaaacaat cagagcttct gctgaatgct gatgatcttg atgccatgtg	360
ggtttgctc agagaaaatt gtgttattga tgatgctact ggtgctgaaa agatgaatta	420
tgaagatttc tgccatatcg ccacagtcctg cactgagtag attggtcaga aatgcaaacg	480
atttttcagc ccttcaaact tcatgaagtc tgcacggagc acttgcacag attgtttggc	540
taattacaag attatctcag tgttttggtt tgaatttaga gtatacttat gtatgaaata	600
ttgattggta ctcatattata ttatattaat tatattatta tt	

(2) INFORMATION FOR SEQ ID NO:471:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 85 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
 (B) LOCATION: 1..85
 (D) OTHER INFORMATION: / Ceres Seq. ID 1482300

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:471:

Lys	Phe	Phe	Ser	Ser	Arg	Arg	Ala	Pro	Asp	Pro	Tyr	Leu	Cys	Ala	Arg
1			5					10					15		
Cys	Ile	Gly	Ala	Arg	Arg	Glu	Gly	Val	Pro	Pro	Ala	Ser	Ser	Met	Leu
			20					25					30		
Trp	Val	Arg	Asn	Ile	Arg	Arg	Phe	Val	Asp	Thr	Gly	Ala	Gly	Leu	Gly
			35				40					45			
Ser	Glu	Ala	Ile	Met	Glu	Leu	Glu	Thr	Lys	Arg	Ile	Leu	Leu	Glu	Ile
			50			55				60					
Phe	Lys	Glu	Arg	Gln	Xaa	Glu	Glu	Cys	Arg	Gly	Trp	Phe	His	Pro	Lys
65				70					75					80	
Phe	Leu	Gln	Glu	Thr											

(2) INFORMATION FOR SEQ ID NO:472:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 152 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
 (B) LOCATION: 1..152
 (D) OTHER INFORMATION: / Ceres Seq. ID 1482301

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:472:

Asn	Phe	Ser	Pro	Ala	Ala	Ala	Leu	Leu	Ile	Leu	Ile	Ser	Ala	Arg	Ala
1			5					10					15		
Ala	Ser	Ala	Pro	Ala	Gly	Arg	Glu	Ser	Arg	Pro	Pro	Arg	Pro	Cys	Cys
			20					25					30		
Gly	Ser	Ala	Ile	Ser	Ala	Ala	Ser	Thr	Arg	Ala	Pro	Ala	Ser	Asp	

```

      35              40              45
Pro Arg Pro Ser Trp Asn Trp Arg Leu Lys Gly Tyr Cys Leu Arg Phe
      50              55              60
Ser Arg Ser Gly Ser Xaa Lys Ser Ala Glu Ala Gly Ser Ile Pro Ser
65              70              75              80
Phe Tyr Lys Lys Pro Glu Glu Gly Ser Ile Ser Ser Arg Val Gln Arg
      85              90              95
Leu Ala Lys Tyr Arg Phe Leu Lys Lys Gln Ser Glu Leu Leu Leu Asn
      100             105             110
Ala Asp Asp Leu Asp Ala Met Trp Val Cys Leu Arg Glu Asn Cys Val
      115             120             125
Ile Asp Asp Ala Thr Gly Ala Glu Lys Met Asn Tyr Glu Asp Phe Cys
      130             135             140
His Ile Ala Thr Val Cys Thr Glu
145              150
```

(2) INFORMATION FOR SEQ ID NO:473:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 607 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..607
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482302

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:473:

```

aagttcgatg gaatagctgg atccgtgacc gtgcaaccac ctcaaccgca tttcaggggc      60
aaggtcttca tcccgaatcc tgggcctcgt ctcccatctc gtgccctaca taaaatccag      120
acgccccgct ttaccccggt cacgacggct ccgcgcttgg ggctgttgaa gccgcctcca      180
gtgttgggcg aggtgccgga tgtgcgttgg tcagtctgcg gccgccactg tctccacact      240
gggtgcgtgc ttggtttggt gacattgaag attatctctg gggcacgacg atgtcgttca      300
tcgctgctgc agcttggttg cctccgatgg ctgcttgta cagcgccggg cctcacgcct      360
cgggacaact tgcgcgctgt catcatccaa ctccattccc gtcgcagtca cgcgtgaggt      420
cacgtgtaat ttgaagatca ttctcgtacc tggaatctaa agtcccaggt caagaacagg      480
taccacagga tgagggggcat ggaggtggct gcgatgtgtt cgtgagargt ctaggtcgtc      540
gtctcccgat caactttggg ttgctggatc gttgtctcct tacgatgtat ttatttattt      600
yatatag
```

(2) INFORMATION FOR SEQ ID NO:474:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 138 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..138
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482303

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:474:

```

Lys Phe Asp Gly Ile Ala Gly Ser Val Thr Val Gln Pro Pro Gln Pro
1              5              10              15
His Phe Arg Gly Lys Val Phe Ile Pro Asn Pro Arg Pro Arg Leu Pro
      20              25              30
Ser Arg Ala Leu His Lys Ile Gln Thr Pro Arg Phe Thr Pro Phe Thr
      35              40              45
Thr Ala Pro Arg Leu Gly Leu Leu Lys Pro Pro Pro Val Leu Gly Glu
      50              55              60
Val Pro Asp Val Arg Trp Ser Val Cys Gly Arg His Cys Leu His Thr
```


(2) INFORMATION FOR SEO ID NO:475:

(A) LENGTH: 546 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ix) FEATURE:

- (A) NAME/KEY: -
(B) LOCATION: 1..546
(D) OTHER INFORMATION: / Ceres Seq. ID 1482307

aaaccctagc	cattttcttt	tcttatcccc	agccgccaca	aagacagccg	gccgcctggg	60
aaactttttt	tctttttctt	ttcctggccg	accgcacctc	ccacttctct	ctattttttt	120
tccttgaccc	atggcttggc	aaacggatct	aggctggctt	cctctctttc	tgttcttctc	180
ctgctctgtt	ctttgcttct	ttcttcccca	ccgaacagct	caaaccggtg	agctagccgc	240
tgctgagcg	cgcagcsggc	tcacatatcc	gaggatgacc	ggattaatgt	tcaggggcat	300
gcggggttta	ccgcgcgtgt	gcgcgtcccc	tcggtagggc	gcgccgccgt	cgggtcaaacg	360
ccgacagatg	tccctgttga	cagcagagaa	aaggggtcga	cctcttggtc	gacttctgtg	420
actgtgcacg	cccagatctt	atctccctta	cctctttgac	tcgatctcgt	tctctgtgat	480
gtatgagatg	gatctaattg	aatattcggg	atcgagttat	gagttgatag	aacgtagatt	540
ttctgc						

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 81 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear

(ix) FEATURE:

- ```
{A} NAME/KEY: peptide
{B} LOCATION: 1..81
{D} OTHER INFORMATION: / Ceres Seq. ID 1482308
```

[illegible]

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 49 amino acids  
(B) TYPE: amino acid

(C) STRANDEDNESS:  
(D) TOPOLOGY: linear  
(ii) MOLECULE TYPE: peptide  
(ix) FEATURE:  
(A) NAME/KEY: peptide  
(B) LOCATION: 1..49  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482309  
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:477:  
Thr Leu Ala Ile Phe Phe Ser Tyr Pro Gln Pro Pro Gln Arg Gln Pro  
1 5 10 15  
Ala Ala Trp Glu Leu Phe Phe Leu Phe Phe Phe Leu Ala Asp Arg Thr  
20 25 30  
Ser His Phe Ser Leu Phe Phe Phe Leu Ala Pro Trp Leu Gly Lys Arg  
35 40 45  
Ile

(2) INFORMATION FOR SEQ ID NO:478:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 742 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear  
(ii) MOLECULE TYPE: DNA (genomic)  
(ix) FEATURE:

(A) NAME/KEY: -  
(B) LOCATION: 1..742  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482322

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:478:

```
atttgctcac caccgccagc accgcagcta gtccattgca ttgacgcctc gatcagggct 60
agcgacggac gaaagaaagc tctgcatgca gcgcctcgcc gccgccgtcg tccccagcct 120
ggtgccgccg ctctacctgt ccatggccgc ctccgccgct gccggctgtt ttccagcaga 180
agcagccggc agtagtagcc ggacgacgac gtcgacgccg acgccgacgc ggcggccatt 240
attagcgcas cgccgtggcg ggtggtgcta ctgacgctcc tgctgctggc gccgagctgc 300
tgccaggcga cgcgaggcat gcagccgttc aggggcaagc cgctgcggcc aggcaccgcc 360
aaccatttcc tggggttctt gccgcgggga ccggcgccctc cgctccggccc ctgcgggcag 420
cacaactcca tcggagcgca ggatcaaagc catccctgac ggcgaccgca ggactgaagc 480
gtggaagaag cagggccgcc gtcgtgtoga tgttcggatc cgaggagtaa gatctccacc 540
aatcaagag agttcgcata accatggatt aggttccttg tcaaaagggtt aagctcgtag 600
tattgattat ttagctagtt tcgtagcact agcagcaata gatgtatact cggagaggga 660
acgaagaaaa ggcacgttct ttgtaggacg atgtacatga ggctatatatt tttttgttgg 720
ggatgggtgt ggtggcgctc cg
```

(2) INFORMATION FOR SEQ ID NO:479:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 64 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS:  
(D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: peptide  
(ix) FEATURE:

(A) NAME/KEY: peptide  
(B) LOCATION: 1..64  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482323

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:479:

Ile Cys Ser Pro Arg Pro Ala Pro Gln Leu Val His Cys Ile Asp Ala  
1 5 10 15  
Ser Ile Arg Ala Ser Asp Gly Arg Lys Lys Ala Leu His Ala Ala Pro  
20 25 30  
Arg Arg Arg Arg Arg Pro Gln Pro Gly Ala Ala Ala Leu Pro Val His  
35 40 45

Gly Arg Leu Arg Arg Cys Arg Leu Phe Ser Ser Arg Ser Ser Arg Gln  
50 55 60

(2) INFORMATION FOR SEQ ID NO:480:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 62 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..62
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482324

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:480:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Gln | Arg | Leu | Ala | Ala | Ala | Val | Val | Pro | Ser | Leu | Val | Pro | Pro | Leu |
| 1   |     |     | 5   |     |     |     | 10  |     |     |     | 15  |     |     |     |     |
| Tyr | Leu | Ser | Met | Ala | Ala | Ser | Ala | Ala | Gly | Cys | Phe | Pro | Ala | Glu |     |
|     |     |     | 20  |     |     |     | 25  |     |     |     | 30  |     |     |     |     |
| Ala | Ala | Gly | Ser | Ser | Ser | Arg | Thr | Thr | Ser | Thr | Pro | Thr | Pro | Thr |     |
|     |     |     | 35  |     |     |     | 40  |     |     |     | 45  |     |     |     |     |
| Arg | Arg | Pro | Leu | Leu | Ala | Xaa | Arg | Arg | Gly | Gly | Trp | Cys | Tyr |     |     |
|     |     |     | 50  |     |     |     | 55  |     |     |     | 60  |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:481:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 46 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..46
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482325

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:481:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Gln | Pro | Phe | Arg | Gly | Lys | Pro | Leu | Arg | Pro | Gly | Thr | Ala | Asn | His |
| 1   |     |     | 5   |     |     |     | 10  |     |     |     | 15  |     |     |     |     |
| Phe | Leu | Gly | Phe | Leu | Pro | Arg | Gly | Pro | Ala | Pro | Pro | Ser | Gly | Pro | Ser |
|     |     |     | 20  |     |     |     | 25  |     |     |     | 30  |     |     |     |     |
| Arg | Gln | His | Asn | Ser | Ile | Gly | Ala | Gln | Asp | Gln | Ser | His | Pro |     |     |
|     |     |     | 35  |     |     |     | 40  |     |     |     | 45  |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:482:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 587 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..587
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482334

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:482:

|             |            |             |            |            |             |     |
|-------------|------------|-------------|------------|------------|-------------|-----|
| acgaccacaca | ccgctgccgc | caccgctgcc  | gacgtakbca | cggccgctcc | ccgaccacaca | 60  |
| cctcgcacta  | tgtsscccc  | accgccgccg  | cctcccctct | agatcctcaa | tgcaactgcta | 120 |
| ggtccegtcc  | acagccgctc | actgccgccca | cctctaacgc | gcctaggacc | atcgccacct  | 180 |
| ccacatctag  | cttctggagt | cgagatccat  | ggtcgactcc | cctggaggag | cccgatctg   | 240 |
| gcctaccggg  | cacggtctcg | agcttttaggc | gtctcagcag | gcagcgcggt | gtactccgtc  | 300 |

```
atcgcccaat ggagcagccg accaagcaag aactctatcc gtgctcgtgc cgaggcgctg 360
tctctactct ctactccatc tgttctgttc ccgcgcctgc gcgtcgtcct ctacggatcc 420
gtccaccgcc gcgccasacc atgtgaactg agacacgcct cmacctatgc atccaagaca 480
casctctgca tctgcgtccg tgcactggct acactggatc gattacggag tggagggtgt 540
tctttacaaa gaaagcttgt accttaaaac aggaggatac aagaagt
```

(2) INFORMATION FOR SEQ ID NO:483:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 147 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..147

(D) OTHER INFORMATION: / Ceres Seq. ID 1482335

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:483:

```
Asp Pro His Arg Cys Arg His Arg Cys Arg Arg Xaa His Gly Arg Ser
1 5 10 15
Pro Thr His Thr Ser His Tyr Xaa Xaa Pro Thr Ala Ala Ala Ser Pro
20 25 30
Leu Asp Pro Gln Cys Thr Ala Arg Ser Arg Pro Gln Pro Leu Thr Ala
35 40 45
Ala Thr Ser Asn Ala Pro Arg Thr Ile Ala Thr Ser Thr Ser Ser Phe
50 55 60
Trp Ser Arg Asp Pro Trp Ser Thr Pro Leu Glu Glu Pro Gly Ser Gly
65 70 75 80
Leu Pro Gly Thr Val Ser Ser Phe Arg Arg Leu Ser Arg Gln Arg Gly
85 90 95
Val Leu Arg His Arg Pro Met Glu Gln Pro Thr Lys Gln Glu Leu Tyr
100 105 110
Pro Cys Ser Cys Arg Gly Ala Val Ser Thr Leu Tyr Ser Ile Cys Ser
115 120 125
Val Pro Ala Pro Ala Arg Arg Pro Leu Arg Ile Arg Pro Pro Pro Arg
130 135 140
Xaa Thr Met
145
```

(2) INFORMATION FOR SEQ ID NO:484:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 543 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..543

(D) OTHER INFORMATION: / Ceres Seq. ID 1482336

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:484:

```
gcttgatgaag acagaaaacg attcctctcc cctccctccc agctctggac gcgtagmgt 60
cgggggcggc cgcctcctcg gctcctcgcc tcacctcccg gtccatcctc gccgctctgc 120
gcgtgcctca cctcgacacc agccttcctt cgtgacacga ctgcaacctc gctgacggag 180
agtacgtcct cgtgccggag caaggtattg ctcaggaggt agccccaga tccagcacca 240
gagcctgcac cagaggatct gcctgccact gctttggaag gttctttgga ggacatggtt 300
gctggagtga cttggccgtc catcttgcca cggggttga cagtcgagtg ggatcctgcc 360
tcggctgagg aggagcatga ggagtgatgg gacaggcttc cccatccctc catttaatta 420
tcgttagttt tattgccgct gcacttcgaa caatgatggc aacttttgaa aaactccgat 480
ggtgatgtaa taatttagta ctccttgatg tatgatatta tgtcttattg tatttgcctc 540
gtg
```

(2) INFORMATION FOR SEQ ID NO:485:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 73 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..73
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482337

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:485:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Ala | Cys | Glu | Asp | Arg | Lys | Arg | Phe | Leu | Ser | Pro | Pro | Ser | Gln | Leu | Trp |
| 1   |     |     |     | 5   |     |     |     | 10  |     |     |     |     |     | 15  |     |
| Thr | Arg | Xaa | Ala | Arg | Arg | Arg | Pro | His | Pro | Arg | Leu | Leu | Ala | Ser | Pro |
|     |     |     | 20  |     |     |     | 25  |     |     |     |     |     | 30  |     |     |
| Pro | Gly | Pro | Ser | Ser | Pro | Leu | Cys | Ala | Cys | Leu | Thr | Ser | Thr | Pro | Ala |
|     |     |     | 35  |     |     |     | 40  |     |     |     |     | 45  |     |     |     |
| Phe | Pro | Arg | Asp | Thr | Thr | Ala | Thr | Ser | Leu | Thr | Glu | Ser | Thr | Ser | Ser |
|     |     |     | 50  |     |     | 55  |     |     |     |     | 60  |     |     |     |     |
| Cys | Arg | Ser | Lys | Val | Leu | Arg | Arg |     |     |     |     |     |     |     |     |
| 65  |     |     |     | 70  |     |     |     |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:486:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 57 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..57
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482338

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:486:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Leu | Val | Lys | Thr | Glu | Asn | Asp | Ser | Ser | Pro | Leu | Pro | Pro | Ser | Ser | Gly |
| 1   |     |     |     | 5   |     |     |     |     | 10  |     |     |     |     | 15  |     |
| Arg | Val | Xaa | Leu | Gly | Gly | Gly | Arg | Ile | Leu | Gly | Ser | Ser | Pro | His | Leu |
|     |     |     | 20  |     |     |     | 25  |     |     |     |     |     | 30  |     |     |
| Pro | Val | His | Pro | Arg | Arg | Ser | Ala | Arg | Ala | Ser | Pro | Arg | His | Gln | Pro |
|     |     |     | 35  |     |     |     | 40  |     |     |     |     |     | 45  |     |     |
| Ser | Leu | Val | Thr | Arg | Leu | Gln | Pro | Arg |     |     |     |     |     |     |     |
|     |     |     | 50  |     |     | 55  |     |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:487:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 633 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..633
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482339

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:487:

|             |            |            |            |             |            |     |
|-------------|------------|------------|------------|-------------|------------|-----|
| acgccgagca  | ctccttctcc | tcctcctctg | tcggcggctg | tgaggagacgt | acacggcgat | 60  |
| taggaggcac  | gtcgtccacc | agtctcctcg | cagggatgtc | gaagagcacg  | gaaatcgcat | 120 |
| ataaagcaat  | catcttgatg | caggatcatg | ccaagcatat | ctatcgtatt  | tgcaatgaga | 180 |
| agctaattatt | gggtaaagga | ttgactgcac | ttgaggtcaa | agaacttcgt  | gaagcacttg | 240 |
| aattcgccgc  | cgaaggattg | gaccagggct | cccttttttg | ccaagaggaa  | ttggatgcaa | 300 |

```
ctgttaagga ggaacaattg gagcatgacg agaaggtggc ttcacagatg attgaaagcc 360
cacttccttc tcctgattcg gactgcttcc tatcccttga agagcacatt gagaagtttt 420
ggggcggttg ttacaactcg gaccagatgc ctagctactc cgactaggct cagagtttat 480
ggtgctgtga aattctagat gtttgggtgt aatgggtatgt tggatgtgta tgtgaactgt 540
aattctggat gtgtggatgt aatgggtgaac tgactgaatg gtgtcttgtg taatgggtatt 600
ttggatgtct atgtgaactc tagctctggg ttt
```

(2) INFORMATION FOR SEQ ID NO:488:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 154 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..154
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482340

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:488:

```
Ala Glu His Ser Phe Ser Ser Ser Ser Val Gly Gly Arg Gly Arg Arg
1 5 10 15
Thr Arg Arg Leu Gly Gly Thr Ser Ser Thr Ser Leu Leu Ala Gly Met
20 25 30
Ser Lys Ser Thr Glu Ile Ala Asp Lys Ala Ile Ile Leu Met Gln Asp
35 40 45
His Ala Lys His Ile Tyr Arg Ile Cys Asn Glu Lys Leu Ile Leu Gly
50 55 60
Lys Gly Leu Thr Ala Phe Glu Val Lys Glu Leu Arg Glu Ala Leu Glu
65 70 75 80
Phe Ala Ala Glu Gly Leu Asp Gln Gly Ser Leu Phe Cys Gln Glu Glu
85 90 95
Leu Asp Ala Thr Val Lys Glu Glu Gln Leu Glu His Asp Glu Lys Val
100 105 110
Ala Ser Gln Met Ile Glu Ser Pro Leu Pro Ser Pro Asp Ser Asp Cys
115 120 125
Phe Leu Ser Leu Glu Glu His Ile Glu Lys Phe Trp Gly Val Asp Tyr
130 135 140
Asn Ser Asp Gln Met Pro Ser Tyr Ser Asp
145 150
```

(2) INFORMATION FOR SEQ ID NO:489:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 123 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..123
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482341

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:489:

```
Met Ser Lys Ser Thr Glu Ile Ala Asp Lys Ala Ile Ile Leu Met Gln
1 5 10 15
Asp His Ala Lys His Ile Tyr Arg Ile Cys Asn Glu Lys Leu Ile Leu
20 25 30
Gly Lys Gly Leu Thr Ala Phe Glu Val Lys Glu Leu Arg Glu Ala Leu
35 40 45
Glu Phe Ala Ala Glu Gly Leu Asp Gln Gly Ser Leu Phe Cys Gln Glu
50 55 60
Glu Leu Asp Ala Thr Val Lys Glu Glu Gln Leu Glu His Asp Glu Lys
```

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |  |  |    |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|--|--|----|
| 65  |     |     |     |     |     | 70  |     |     |     |     |     | 75  |     |     |     |  |  | 80 |
| Val | Ala | Ser | Gln | Met | Ile | Glu | Ser | Pro | Leu | Pro | Ser | Pro | Asp | Ser | Asp |  |  |    |
|     |     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |  |  |    |
| Cys | Phe | Leu | Ser | Leu | Glu | Glu | His | Ile | Glu | Lys | Phe | Trp | Gly | Val | Asp |  |  |    |
|     |     |     | 100 |     |     |     |     | 105 |     |     |     |     | 110 |     |     |  |  |    |
| Tyr | Asn | Ser | Asp | Gln | Met | Pro | Ser | Tyr | Ser | Asp |     |     |     |     |     |  |  |    |
|     |     | 115 |     |     |     |     | 120 |     |     |     |     |     |     |     |     |  |  |    |

(2) INFORMATION FOR SEQ ID NO:490:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 109 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS:  
(D) TOPOLOGY: linear

## (ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide  
(B) LOCATION: 1..109  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482342

(xi) SEQUENCE DESCRIPTION: SEO ID NO:490:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Gln | Asp | His | Ala | Lys | His | Ile | Tyr | Arg | Ile | Cys | Asn | Glu | Lys | Leu |
| 1   |     |     |     | 5   |     |     |     |     | 10  |     |     |     |     | 15  |     |
| Ile | Leu | Gly | Lys | Gly | Leu | Thr | Ala | Phe | Glu | Val | Lys | Glu | Leu | Arg | Glu |
|     |     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |
| Ala | Leu | Glu | Phe | Ala | Ala | Glu | Gly | Leu | Asp | Gln | Gly | Ser | Leu | Phe | Cys |
|     |     |     | 35  |     |     |     | 40  |     |     |     |     | 45  |     |     |     |
| Gln | Glu | Glu | Leu | Asp | Ala | Thr | Val | Lys | Glu | Glu | Gln | Leu | Glu | His | Asp |
|     | 50  |     |     |     |     | 55  |     |     |     |     | 60  |     |     |     |     |
| Glu | Lys | Val | Ala | Ser | Gln | Met | Ile | Glu | Ser | Pro | Leu | Pro | Ser | Pro | Asp |
| 65  |     |     |     |     | 70  |     |     |     |     | 75  |     |     |     |     | 80  |
| Ser | Asp | Cys | Phe | Leu | Ser | Leu | Glu | Glu | His | Ile | Glu | Lys | Phe | Trp | Gly |
|     |     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |
| Val | Asp | Tyr | Asn | Ser | Asp | Gln | Met | Pro | Ser | Tyr | Ser | Asp |     |     |     |
|     |     |     | 100 |     |     |     |     | 105 |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:491:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 827 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -  
(B) LOCATION: 1..827  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482346

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:491:

|             |            |             |             |            |             |     |
|-------------|------------|-------------|-------------|------------|-------------|-----|
| cctaatacgaa | aaatcgaaaa | cccaccgcac  | cctttcatca  | gcctgcctgt | ccactgttgg  | 60  |
| cttggtgact  | tcttcgcctc | cgctccgcctc | ccctccgcctc | ccgaacggtc | gatctttgca  | 120 |
| tggcagcagc  | agctggctcc | aagggggcggg | cgatcgctgg  | aagcttcgtc | agccgcgtcc  | 180 |
| tcgccggcaa  | ggccgcctcg | ccgaggaggg  | ccgtgcacgc  | ctcggcgta  | gacaagaacc  | 240 |
| tggaggacca  | ggtgcgccc  | gcgttcgtgc  | cggacgatgt  | gatcggcagc | gccggngagc  | 300 |
| cccgacaagt  | actggagccc | ccaccccaag  | accggcgctc  | tccggccggc | ggcgggtggac | 360 |
| cccaagctgg  | ccgctggtgg | cgcgccggga  | cgcggcgcg   | gawtgcgtca | ggaggcacgg  | 420 |
| tgctggacca  | gaaggtgtgg | ttccgcccgc  | tcgaggacgt  | cgagaagccg | ccccccgcgc  | 480 |
| cgtgagccgc  | gcggcgctgc | taggccagcc  | cacactgctg  | ctcgctcata | aaaagggcgg  | 540 |
| cgggagagcc  | tggcagtggc | aggcactctg  | ctcgtgctcg  | gccgggctgg | gctccctgct  | 600 |
| tataatcactg | caatattata | ctactagtag  | tggtgcttga  | tagcagtgtg | tggctgtgct  | 660 |
| aataaccagta | taatactggt | tctactataa  | tacagtcgta  | tcaggcatgg | cgtgcatacag | 720 |
| gactggttgt  | gatagtagca | acgtgatgct  | cgtgcctgta  | ataagaacaa | gcaggcgatg  | 780 |
| tgtgcctgtg  | atgtaccggt | gtcgtcagtg  | ttataaqtac  | ttggggc    |             |     |

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 212 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ix) FEATURE:

(B) LOCATION: 1..212

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:492:

(2) INFORMATION FOR SEQ ID NO:493:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 121 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ix) FEATURE:

(B) LOCATION: 1..121

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:493:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Ala | Ala | Ala | Ala | Gly | Ser | Lys | Gly | Arg | Ala | Ile | Ala | Gly | Ser | Phe |
| 1   |     |     |     | 5   |     |     |     |     | 10  |     |     |     |     | 15  |     |
| Val | Ser | Arg | Val | Leu | Ala | Gly | Lys | Ala | Ala | Ser | Pro | Arg | Arg | Ala | Val |
|     |     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |
| His | Ala | Ser | Ala | Tyr | Asp | Lys | Asn | Leu | Glu | Asp | Gln | Val | Arg | Pro | Ala |
|     |     |     | 35  |     |     |     | 40  |     |     |     |     |     | 45  |     |     |
| Phe | Val | Pro | Asp | Asp | Val | Ile | Gly | Ser | Ala | Xaa | Glu | Pro | Arg | Gln | Val |



```

50 55 60
Leu Glu Pro Pro Pro Gln Asp Arg Arg Leu Arg Pro Gly Gly Gly Gly
65 70 75 80
Pro Gln Ala Gly Arg Trp Trp Arg Arg Arg Thr Pro Ala Arg Xaa Ala
85 90 95
Ala Gly Gly Thr Val Leu Asp Gln Lys Val Trp Phe Arg Pro Leu Glu
100 105 110
Asp Val Glu Lys Pro Pro Pro Ala Ala
115 120
```

(2) INFORMATION FOR SEQ ID NO:494:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 767 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..767
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482349

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:494:

```

awrtcgcgat cgcgtccttc tcttcccaac ttcctcgcg tcatctcttc accccacccg 60
ccccaaaacg ccaaatctaa cagcaaaggt ccggaacctt ctagccgcac ctagggtttg 120
gattggcgcc gagcatggcg tacgtcgacc acgccttctc catctccgac gaggacgacc 180
tcgtcggmgy cgccatgggg gggccgcgcg gggcgcmcg tgaaggagatc gccttcgccc 240
ccgcgctgct cgmcttcggg gcgbtcggtt ccatcagggt gcctgctaata ggctgtcaac 300
cgcgtcggag gggaccgcgc gcacggaatt ttcttcatga tgttgggcat tgtaatgttc 360
atccctgggt tctactacac aaggatcgcc tactatgctt acaaagggtta caagggtttc 420
tctttttcga acatcccacc gatctgaagg agtgtgctgc ctgcctggct ggcatgaag 480
tggtgtcgct ggtttaagag tttgtcgact ctgtogaatg gctctgtaga cacccttgtt 540
ctacatcttt ctgtggccac attctctttg aacactctag aatgaactgg tggatgtgta 600
cagataaatg cagccatagt tgtgtcccat cgctgtttgg ccgattggaa ggttgtttgt 660
tgtgctagtg tgaccatggt caactgatac gcattgctac ttgtgcatta ctatcgtttt 720
tgtcagggac cttaaatacat tatatgggaa taagatctcg tcgttcc
```

(2) INFORMATION FOR SEQ ID NO:495:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 95 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..95
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482350

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:495:

```

Xaa Ser His Arg Val Leu Leu Phe Pro Thr Ser Ser Arg Ser Phe Leu
1 5 10 15
His Pro Thr Arg Pro Gln Thr Pro Asn Leu Thr Ala Lys Val Arg Asn
20 25 30
Leu Leu Ala Ala Pro Arg Val Trp Ile Gly Ala Glu His Gly Val Arg
35 40 45
Arg Pro Arg Leu Leu His Leu Arg Arg Gly Arg Pro Arg Arg Xaa Arg
50 55 60
His Gly Gly Pro Ala Arg Gly Xaa Arg Glu Gly Asp Arg Leu Arg Arg
65 70 75 80
Arg Ala Ala Xaa Leu Arg Gly Xaa Arg Tyr His Gln Val Ala Cys
85 90 95
```

(2) INFORMATION FOR SEQ ID NO:496:

(i) SEQUENCE CHARACTERISTICS:  
    (A) LENGTH: 120 amino acids  
    (B) TYPE: amino acid  
    (C) STRANDEDNESS:  
    (D) TOPOLOGY: linear  
(ii) MOLECULE TYPE: peptide  
(ix) FEATURE:  
    (A) NAME/KEY: peptide  
    (B) LOCATION: 1..120  
    (D) OTHER INFORMATION: / Ceres Seq. ID 1482351  
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:496:  
Met Ala Tyr Val Asp His Ala Phe Ser Ile Ser Asp Glu Asp Asp Leu  
1                  5                  10                  15  
Val Xaa Gly Ala Met Gly Gly Pro Arg Gly Ala Xaa Val Lys Glu Ile  
                  20                  25                  30  
Ala Phe Ala Ala Ala Leu Leu Xaa Phe Gly Ala Xaa Gly Thr Ile Arg  
                  35                  40                  45  
Trp Pro Ala Asn Gly Cys Gln Pro Arg Arg Arg Gly Pro Arg Ala Arg  
50                  55                  60  
Asn Phe Leu His Asp Val Gly His Cys Asn Val His Pro Trp Val Leu  
65                  70                  75                  80  
Leu His Lys Asp Arg Leu Leu Cys Leu Gln Arg Leu Gln Gly Phe Leu  
                  85                  90                  95  
Phe Phe Glu His Pro Thr Asp Leu Lys Glu Cys Ala Ala Cys Leu Ala  
100                  105                  110  
Gly His Glu Val Val Ser Leu Val  
115                  120

(2) INFORMATION FOR SEQ ID NO:497:  
    (i) SEQUENCE CHARACTERISTICS:  
        (A) LENGTH: 100 amino acids  
        (B) TYPE: amino acid  
        (C) STRANDEDNESS:  
        (D) TOPOLOGY: linear  
    (ii) MOLECULE TYPE: peptide  
    (ix) FEATURE:  
        (A) NAME/KEY: peptide  
        (B) LOCATION: 1..100  
        (D) OTHER INFORMATION: / Ceres Seq. ID 1482352  
    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:497:  
Met Gly Gly Pro Arg Gly Ala Xaa Val Lys Glu Ile Ala Phe Ala Ala  
1                  5                  10                  15  
Ala Leu Leu Xaa Phe Gly Ala Xaa Gly Thr Ile Arg Trp Pro Ala Asn  
                  20                  25                  30  
Gly Cys Gln Pro Arg Arg Arg Gly Pro Arg Ala Arg Asn Phe Leu His  
35                  40                  45  
Asp Val Gly His Cys Asn Val His Pro Trp Val Leu Leu His Lys Asp  
50                  55                  60  
Arg Leu Leu Cys Leu Gln Arg Leu Gln Gly Phe Leu Phe Phe Glu His  
65                  70                  75                  80  
Pro Thr Asp Leu Lys Glu Cys Ala Ala Cys Leu Ala Gly His Glu Val  
                  85                  90                  95  
Val Ser Leu Val  
100

(2) INFORMATION FOR SEQ ID NO:498:  
    (i) SEQUENCE CHARACTERISTICS:  
        (A) LENGTH: 1072 base pairs  
        (B) TYPE: nucleic acid  
        (C) STRANDEDNESS: single  
        (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..1072

(D) OTHER INFORMATION: / Ceres Seq. ID 1482353

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:498:

|            |            |            |             |            |             |      |
|------------|------------|------------|-------------|------------|-------------|------|
| acatcacaaa | ccgaaaaarg | ccgcgacgag | ccgacgatct  | ctactgcccc | cttccggcct  | 60   |
| tcggcgaccg | tgacgagcaa | cgacgacgac | ggcgacgatg  | gccgcttcct | ccctctgcca  | 120  |
| cgggcacttg | ctcctgtttc | tcctcgtgtc | cgtcacatcg  | gcctgcctcg | gtaccgcggc  | 180  |
| asscantcaa | gccgggtctg | gagagggcta | cacgatcgcc  | ggccgcgtca | agatcgatgg  | 240  |
| catgagttag | aagggctatg | gtcttccagc | caagacatca  | aacacaaaag | tgatacttaa  | 300  |
| tggcggccaa | agggttacat | ttgccaggcc | agacggctac  | tttgatttcc | acaacgtgcc  | 360  |
| agctggaact | catctgattg | aggtctcctc | aattggttac  | ttctttttcc | ctgtccgagt  | 420  |
| tgatataagt | gcaaggaatc | ctggatatat | tcaagcagca  | ttgactgaaa | ccagaagagt  | 480  |
| tctgaatgag | cttgttctgg | aacctctgaa | agaagagcag  | tactttgagg | ttagggagcc  | 540  |
| gttctccgtc | atgtcacttt | tgaagagccc | catggggtta  | atggttgggt | ttatgggtctt | 600  |
| aatggctctc | gtgatgcccc | agatgatgga | gaacatagat  | cccaggagga | tgaagcaagc  | 660  |
| tcaagwacaa | atgaggaaca | accctgtatc | attctctggc  | ttgctcgcca | gagcgcaggg  | 720  |
| ctagagaagt | agactgtaga | catgaggata | ctgcaaagggt | caaacattct | agaatgtgag  | 780  |
| taagagcact | attaaagtgc | ttggcacgtc | actcactcgg  | ggcaatttcc | tggggataag  | 840  |
| aaggaaatcc | tttccccttg | tttttaccgt | attttagggc  | tagtttgggg | acaccaattt  | 900  |
| tccaaaggat | ttatatattt | ccatgggaaa | atgaactaat  | tttccttggg | aaaatgaaaa  | 960  |
| tctcttgtaa | aattgggggt | ccaaactagy | ccttaagtta  | taatttgtct | gcggtgtaga  | 1020 |
| accttctgaa | acctctgagc | tagtgatgcg | tcagattgag  | atattttgtt | cg          |      |

(2) INFORMATION FOR SEQ ID NO:499:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 208 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..208

(D) OTHER INFORMATION: / Ceres Seq. ID 1482354

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:499:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Ala | Ala | Ser | Ser | Leu | Cys | His | Gly | His | Leu | Leu | Leu | Phe | Leu | Leu |
| 1   |     |     |     | 5   |     |     |     |     | 10  |     |     |     |     | 15  |     |
| Val | Ser | Val | Thr | Ser | Ala | Cys | Leu | Gly | Thr | Ala | Ala | Xaa | Xaa | Gln | Ala |
|     |     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |
| Gly | Ser | Gly | Glu | Gly | Tyr | Thr | Ile | Ala | Gly | Arg | Val | Lys | Ile | Asp | Gly |
|     |     | 35  |     |     |     | 40  |     |     |     |     | 45  |     |     |     |     |
| Met | Ser | Glu | Lys | Gly | Tyr | Gly | Leu | Pro | Ala | Lys | Thr | Ser | Asn | Thr | Lys |
|     | 50  |     |     |     |     | 55  |     |     |     |     | 60  |     |     |     |     |
| Val | Ile | Leu | Asn | Gly | Gly | Gln | Arg | Val | Thr | Phe | Ala | Arg | Pro | Asp | Gly |
|     | 65  |     |     |     | 70  |     |     |     |     | 75  |     |     |     | 80  |     |
| Tyr | Phe | Ala | Phe | His | Asn | Val | Pro | Ala | Gly | Thr | His | Leu | Ile | Glu | Val |
|     |     |     | 85  |     |     |     |     |     | 90  |     |     |     |     | 95  |     |
| Ser | Ser | Ile | Gly | Tyr | Phe | Phe | Ser | Pro | Val | Arg | Val | Asp | Ile | Ser | Ala |
|     |     |     | 100 |     |     |     |     | 105 |     |     |     |     | 110 |     |     |
| Arg | Asn | Pro | Gly | Tyr | Ile | Gln | Ala | Ala | Leu | Thr | Glu | Thr | Arg | Arg | Val |
|     |     |     | 115 |     |     |     | 120 |     |     |     |     | 125 |     |     |     |
| Leu | Asn | Glu | Leu | Val | Leu | Glu | Pro | Leu | Lys | Glu | Glu | Gln | Tyr | Phe | Glu |
|     | 130 |     |     |     |     | 135 |     |     |     |     | 140 |     |     |     |     |
| Val | Arg | Glu | Pro | Phe | Ser | Val | Met | Ser | Leu | Leu | Lys | Ser | Pro | Met | Gly |
|     | 145 |     |     |     | 150 |     |     |     |     | 155 |     |     |     | 160 |     |
| Leu | Met | Val | Gly | Phe | Met | Val | Leu | Met | Val | Phe | Val | Met | Pro | Lys | Met |
|     |     |     | 165 |     |     |     |     | 170 |     |     |     |     |     | 175 |     |
| Met | Glu | Asn | Ile | Asp | Pro | Glu | Glu | Met | Lys | Gln | Ala | Gln | Xaa | Gln | Met |

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     | 180 |     | 185 |     | 190 |     |     |     |     |     |     |     |     |     |     |
| Arg | Asn | Asn | Pro | Val | Ser | Phe | Ser | Gly | Leu | Leu | Ala | Arg | Ala | Gln | Gly |
|     | 195 |     |     |     |     |     | 200 |     |     |     |     | 205 |     |     |     |

(2) INFORMATION FOR SEQ ID NO:500:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 160 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..160
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482355

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:500:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Ser | Glu | Lys | Gly | Tyr | Gly | Leu | Pro | Ala | Lys | Thr | Ser | Asn | Thr | Lys |
| 1   |     |     | 5   |     |     |     | 10  |     |     |     |     |     | 15  |     |     |
| Val | Ile | Leu | Asn | Gly | Gly | Gln | Arg | Val | Thr | Phe | Ala | Arg | Pro | Asp | Gly |
|     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |     |
| Tyr | Phe | Ala | Phe | His | Asn | Val | Pro | Ala | Gly | Thr | His | Leu | Ile | Glu | Val |
|     | 35  |     |     |     |     | 40  |     |     |     |     | 45  |     |     |     |     |
| Ser | Ser | Ile | Gly | Tyr | Phe | Phe | Ser | Pro | Val | Arg | Val | Asp | Ile | Ser | Ala |
|     | 50  |     |     |     |     | 55  |     |     |     | 60  |     |     |     |     |     |
| Arg | Asn | Pro | Gly | Tyr | Ile | Gln | Ala | Ala | Leu | Thr | Glu | Thr | Arg | Arg | Val |
|     | 65  |     |     |     | 70  |     |     |     | 75  |     |     |     |     | 80  |     |
| Leu | Asn | Glu | Leu | Val | Leu | Glu | Pro | Leu | Lys | Glu | Glu | Gln | Tyr | Phe | Glu |
|     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |     |
| Val | Arg | Glu | Pro | Phe | Ser | Val | Met | Ser | Leu | Leu | Lys | Ser | Pro | Met | Gly |
|     | 100 |     |     |     |     |     | 105 |     |     |     |     | 110 |     |     |     |
| Leu | Met | Val | Gly | Phe | Met | Val | Leu | Met | Val | Phe | Val | Met | Pro | Lys | Met |
|     | 115 |     |     |     |     | 120 |     |     |     | 125 |     |     |     |     |     |
| Met | Glu | Asn | Ile | Asp | Pro | Glu | Glu | Met | Lys | Gln | Ala | Gln | Xaa | Gln | Met |
|     | 130 |     |     |     |     | 135 |     |     |     | 140 |     |     |     |     |     |
| Arg | Asn | Asn | Pro | Val | Ser | Phe | Ser | Gly | Leu | Leu | Ala | Arg | Ala | Gln | Gly |
|     | 145 |     |     |     | 150 |     |     |     | 155 |     |     |     |     | 160 |     |

(2) INFORMATION FOR SEQ ID NO:501:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 803 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..803
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482356

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:501:

|            |             |             |            |             |            |     |
|------------|-------------|-------------|------------|-------------|------------|-----|
| accaacctca | cctactgttc  | tgggttgaaa  | tcttgcgga  | agtctgccaa  | aacaaaaaac | 60  |
| aaaagtcctt | gcaggtgggt  | tggcaggcta  | aacttgacgt | ttgcgttggc  | aggaagccgt | 120 |
| ggctgctgta | atctaactctg | ctgctgcaat  | ctccgaccg  | tctcccagac  | ttgactgtac | 180 |
| ctgaaaccac | tattgaaaca  | atcgggtgaga | gcgagagaga | aaattaaaga  | gaaacccgac | 240 |
| aaaaaccaac | caaccaagca  | gctctccgtt  | ccatatagcc | gctgcatcag  | atccattcaa | 300 |
| gaactagagc | caagccacca  | acaataaatt  | cctctggccg | gcctgcctca  | tcagctcggt | 360 |
| tcaaaaaaaa | caaaaaaaa   | agaagtcgca  | gcggcagtag | taaaactgcag | tgacatacgg | 420 |
| agcactactg | tactgtactg  | tagtaacata  | ctactactgc | tgctgctcac  | agcaagaaca | 480 |

aggatacgcgta aaaaaagaac caaggcaaaa agctaaggctc ctgtttggga acaaagtttt 540  
tgaaaaccac agtttttgaa atactatact atactttagt tataacaata ccgtagttta 600  
taataccgca gttttgaaaa ctgaggtcca gagctaagtt tagaatgcct taaaacaact 660  
atagtatttg caatacttca gttttgaaaa cagagatttt acctagcttg ccaaacacca 720  
ttatgtatat aatactgcag tatttgagaa tactgcagta ttcttccaaa actgcagaaa 780  
aactttgttc ccaaacaccc cct

(2) INFORMATION FOR SEQ ID NO:502:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 57 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..57
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482357

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:502:

Thr Asn Leu Thr Tyr Cys Ser Gly Leu Lys Ser Cys Gly Lys Ser Ala  
1 5 10 15  
Lys Thr Lys Asn Lys Ser Pro Cys Arg Trp Phe Gly Arg Leu Asn Leu  
20 25 30  
Thr Phe Ala Leu Ala Gly Ser Arg Gly Cys Cys Asn Leu Ile Cys Cys  
35 40 45  
Cys Asn Leu Arg Pro Val Ser Gln Thr  
50 55

(2) INFORMATION FOR SEQ ID NO:503:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 42 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..42
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482358

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:503:

Gln Pro His Leu Leu Phe Trp Val Glu Ile Leu Arg Lys Val Cys Gln  
1 5 10 15  
Asn Lys Lys Gln Lys Ser Leu Gln Val Val Trp Gln Ala Lys Leu Asp  
20 25 30  
Val Cys Val Gly Arg Lys Pro Trp Leu Leu  
35 40

(2) INFORMATION FOR SEQ ID NO:504:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 517 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..517
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482359

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:504:

gtagacgcagc tgcattgtgtr gccggccaat ttacgcgccg ccacatgctc tgctcgccca 60  
tcgctttcga gctttgtgta aatggactag agcgggaaggc atagcatgca taggaatagg 120  
agcaactaac caccggcctc tcgctccctc gctgcgccat aaggctgcga ctgcgagagc 180

```
cagccgcacc cgcaccagtc cataggccgg cctcctctct taccttccca cacccttct 240
cgaccgtacg tagcctagtt gtgcttggtta gccagccaga aggtcgtcgg ccgatgatgg 300
gaggaagaac agtggccccc cgcctcgtcc tcgcgctggt gaccatcatc gccatcgggc 360
gcgggccgarg gggacgaggt gaagtgtggc gggtrctctc cgtrcrgcgg crccgactgc 420
mcggtgctgt acccgthnmm gcscrcrcg ccgtactact actacagcmc tmccccaccc 480
gcgacctacc ccggggagtc ctcgtcatatc taccagc
```

(2) INFORMATION FOR SEQ ID NO:505:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 63 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..63
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482360

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:505:

```
Met Cys Xaa Arg Pro Ile Tyr Ala Pro Pro His Ala Leu Leu Ala His
1 5 10 15
Arg Phe Arg Ala Leu Cys Lys Trp Thr Arg Ala Glu Gly Ile Ala Cys
 20 25 30
Ile Gly Ile Gly Ala Thr Asn His Arg Pro Leu Ala Pro Ser Leu Arg
 35 40 45
His Lys Ala Ala Thr Ala Arg Ala Ser Arg Thr Arg Thr Ser Pro
 50 55 60
```

(2) INFORMATION FOR SEQ ID NO:506:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 74 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..74
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482361

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:506:

```
Met Met Gly Gly Arg Thr Val Ala Pro Pro Leu Val Leu Ala Leu Val
1 5 10 15
Thr Ile Ile Ala Ile Gly Gly Gly Arg Xaa Gly Arg Gly Glu Val Trp
 20 25 30
Arg Xaa Leu Ser Xaa Xaa Arg Xaa Arg Leu Xaa Gly Ala Val Pro Xaa
 35 40 45
Xaa Xaa Pro Xaa Ala Val Leu Leu Leu Gln Xaa Xaa Pro Thr Arg Asp
 50 55 60
Leu Pro Arg Gly Val Leu Val Ile Leu Pro
 65 70
```

(2) INFORMATION FOR SEQ ID NO:507:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 73 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..73
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482362

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:507:

```
Met Gly Gly Arg Thr Val Ala Pro Pro Leu Val Leu Ala Leu Val Thr
1 5 10 15
Ile Ile Ala Ile Gly Gly Gly Arg Xaa Gly Arg Gly Glu Val Trp Arg
20 25 30
Xaa Leu Ser Xaa Xaa Arg Xaa Arg Leu Xaa Gly Ala Val Pro Xaa Xaa
35 40 45
Xaa Pro Xaa Ala Val Leu Leu Leu Gln Xaa Xaa Pro Thr Arg Asp Leu
50 55 60
Pro Arg Gly Val Leu Val Ile Leu Pro
65 70
```

(2) INFORMATION FOR SEQ ID NO:508:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 449 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..449
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482363

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:508:

```
aaagagaaag tttattacga tgtaggtgca tattcaaggc ccgttgatgg atgaactttt 60
gtagtgtggtg tccaaaagggtg tacgtatgtg ggacgggcat aaaaaatatg attttgatct 120
acgtgcttttg ttattggcga acaggcgagt gagtgaagag agaagccatg cctctttcgt 180
gtgaggcaag cgatgaacga gtagatgctg ccattcaaca agggattcag ggtctgcacc 240
tattgttttag atgagatcgg tatcttgtat ctacatcatt gtagagaagt tatttacatg 300
ggccatcgtc gatttcttgt aaacacccaaa taagaagaaa aggcaagcat tgaaatgcac 360
aagtagacca tcgtgccaaa gcctattccc caaaggagca accttgattt ccagatggta 420
tagaacttaa atgtagtgta tgggaatcg
```

(2) INFORMATION FOR SEQ ID NO:509:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 33 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..33
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482364

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:509:

```
Met Asn Phe Cys Ser Cys Gly Pro Lys Val Tyr Val Cys Gly Thr Gly
1 5 10 15
Ile Lys Asn Met Ile Leu Ile Tyr Val Leu Cys Tyr Trp Arg Thr Gly
20 25 30
Glu
```

(2) INFORMATION FOR SEQ ID NO:510:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 32 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..32

(D) OTHER INFORMATION: / Ceres Seq. ID 1482365

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:510:

```
Met Trp Asp Gly His Lys Lys Tyr Asp Phe Asp Leu Arg Ala Leu Leu
1 5 10 15
Leu Ala Asn Arg Arg Val Ser Glu Glu Arg Ser His Ala Ser Phe Val
20 25 30
```

(2) INFORMATION FOR SEQ ID NO:511:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 42 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..42

(D) OTHER INFORMATION: / Ceres Seq. ID 1482366

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:511:

```
Met Leu Pro Phe Asn Lys Gly Phe Arg Val Cys Thr Tyr Cys Leu Asp
1 5 10 15
Glu Ile Gly Ile Leu Tyr Leu His His Cys Arg Glu Val Ile Tyr Met
20 25 30
Gly His Arg Arg Phe Leu Val Asn Thr Lys
35 40
```

(2) INFORMATION FOR SEQ ID NO:512:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 757 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..757

(D) OTHER INFORMATION: / Ceres Seq. ID 1482371

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:512:

```
tgattggttt gatgaacagc tagagaacta cttagatgat gattatcttg tgtttgattg 60
ccctggccag attgaactct tcacacatgt tccagttctg cggaactttg tcgagcacct 120
gaaacgaaaa aatttcaacg tttgcgctgt ttaccttctt gattcacagt ttgtcagcga 180
tgtaacaaaa tacatcagtg gttgcatggc ttctctatct gctatgattc agcttgaact 240
tcctcatatc aacatccttt caaagatgga tctggtctcc aacaaaaaag atgtagaaga 300
gtacctggac ccgaatgcac aggttcttct ttcacagctg aatcggcaga tggcacctcg 360
gtttggcaag ttgaacaagt gtttagctga actggttgat gattacagca tggttaatth 420
cattccactt gatttgagaa aggaaagcag catacaatat gtgctatctt ctatcgacac 480
ctgtatccag tatggggaag atgcagatgt gaaggtcagg gacttcgaag aagacgaaga 540
ctaaccactg gcaactggat ctgtaggagg tgcaaactgg ttgctagcag tcgtgtagtg 600
cggagtgaga ctttgggact gtgtakggtg gcgcaggcat gcaaaaacgt cgtaggatgc 660
tgatgacagc tawctggcct atgtaagacg aactaawgca gatatttggc aagtcctagt 720
aaaatgtgtg wgcrccttga tggmyctrw tctcccc
```

(2) INFORMATION FOR SEQ ID NO:513:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 180 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:



(A) NAME/KEY: peptide  
(B) LOCATION: 1..180  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482372

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:513:

Asp Trp Phe Asp Glu Gln Leu Glu Asn Tyr Leu Asp Asp Asp Tyr Leu  
1 5 10 15  
Val Phe Asp Cys Pro Gly Gln Ile Glu Leu Phe Thr His Val Pro Val  
20 25 30  
Leu Arg Asn Phe Val Glu His Leu Lys Arg Lys Asn Phe Asn Val Cys  
35 40 45  
Ala Val Tyr Leu Leu Asp Ser Gln Phe Val Ser Asp Val Thr Lys Tyr  
50 55 60  
Ile Ser Gly Cys Met Ala Ser Leu Ser Ala Met Ile Gln Leu Glu Leu  
65 70 75 80  
Pro His Ile Asn Ile Leu Ser Lys Met Asp Leu Val Ser Asn Lys Lys  
85 90 95  
Asp Val Glu Glu Tyr Leu Asp Pro Asn Ala Gln Val Leu Leu Ser Gln  
100 105 110  
Leu Asn Arg Gln Met Ala Pro Arg Phe Gly Lys Leu Asn Lys Cys Leu  
115 120 125  
Ala Glu Leu Val Asp Asp Tyr Ser Met Val Asn Phe Ile Pro Leu Asp  
130 135 140  
Leu Arg Lys Glu Ser Ser Ile Gln Tyr Val Leu Ser Ser Ile Asp Thr  
145 150 155 160  
Cys Ile Gln Tyr Gly Glu Asp Ala Asp Val Lys Val Arg Asp Phe Glu  
165 170 175  
Glu Asp Glu Asp  
180

(2) INFORMATION FOR SEQ ID NO:514:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 112 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS:  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide  
(B) LOCATION: 1..112  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482373

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:514:

Met Ala Ser Leu Ser Ala Met Ile Gln Leu Glu Leu Pro His Ile Asn  
1 5 10 15  
Ile Leu Ser Lys Met Asp Leu Val Ser Asn Lys Lys Asp Val Glu Glu  
20 25 30  
Tyr Leu Asp Pro Asn Ala Gln Val Leu Leu Ser Gln Leu Asn Arg Gln  
35 40 45  
Met Ala Pro Arg Phe Gly Lys Leu Asn Lys Cys Leu Ala Glu Leu Val  
50 55 60  
Asp Asp Tyr Ser Met Val Asn Phe Ile Pro Leu Asp Leu Arg Lys Glu  
65 70 75 80  
Ser Ser Ile Gln Tyr Val Leu Ser Ser Ile Asp Thr Cys Ile Gln Tyr  
85 90 95  
Gly Glu Asp Ala Asp Val Lys Val Arg Asp Phe Glu Glu Asp Glu Asp  
100 105 110

(2) INFORMATION FOR SEQ ID NO:515:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 106 amino acids

- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
  - (A) NAME/KEY: peptide
  - (B) LOCATION: 1..106
  - (D) OTHER INFORMATION: / Ceres Seq. ID 1482374
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:515:

```
Met Ile Gln Leu Glu Leu Pro His Ile Asn Ile Leu Ser Lys Met Asp
1 5 10 15
Leu Val Ser Asn Lys Lys Asp Val Glu Glu Tyr Leu Asp Pro Asn Ala
20 25 30
Gln Val Leu Leu Ser Gln Leu Asn Arg Gln Met Ala Pro Arg Phe Gly
35 40 45
Lys Leu Asn Lys Cys Leu Ala Glu Leu Val Asp Asp Tyr Ser Met Val
50 55 60
Asn Phe Ile Pro Leu Asp Leu Arg Lys Glu Ser Ser Ile Gln Tyr Val
65 70 75 80
Leu Ser Ser Ile Asp Thr Cys Ile Gln Tyr Gly Glu Asp Ala Asp Val
85 90 95
Lys Val Arg Asp Phe Glu Glu Asp Glu Asp
100 105
```

- (2) INFORMATION FOR SEQ ID NO:516:
  - (i) SEQUENCE CHARACTERISTICS:
    - (A) LENGTH: 617 base pairs
    - (B) TYPE: nucleic acid
    - (C) STRANDEDNESS: single
    - (D) TOPOLOGY: linear
  - (ii) MOLECULE TYPE: DNA (genomic)
  - (ix) FEATURE:
    - (A) NAME/KEY: -
    - (B) LOCATION: 1..617
    - (D) OTHER INFORMATION: / Ceres Seq. ID 1482375
  - (xi) SEQUENCE DESCRIPTION: SEQ ID NO:516:

```
agaatccccc gtmgacgcgc acggcagagc tccgcacccg caccggccgc cggcggstgg 60
atggggaagc tctccgccct gaagcgggaa gcggtcggag tggactaggc gctcgggtgac 120
ttcctagttt agaagcggta rgtggaggcg atgcggggcg gggcgatgaa ggccctgcgg 180
cgatccagca cctcctcggc gccatcgcca agggtgccgt cttccccgcg gtcttattcg 240
tggatccacc gccggtcgct tctcgttacc tygccggcct cgccggmgmc gtcctctgtg 300
tctgaatcgg cgaatttgcc cgcggagggt tcggattcag cgccagcktc agtgggtggca 360
gcttcctcgt cgccctcgct ggctgcttcg tctccgaaca tggaatggtg gggctatcct 420
gtccggattt ctctctgtgc tgcattgggtc gcttcaaaat ggggaataat tgttgggcta 480
cttgatatatt cccaacaacg attcgcgcac ttattcccaa aattatgctg ttctggtagc 540
acagtggaag tggtagtttg ttcggtacta ttattcttat aagatttgct ttagtctctt 600
agattaaaaa aaagctg
```

- (2) INFORMATION FOR SEQ ID NO:517:
  - (i) SEQUENCE CHARACTERISTICS:
    - (A) LENGTH: 143 amino acids
    - (B) TYPE: amino acid
    - (C) STRANDEDNESS:
    - (D) TOPOLOGY: linear
  - (ii) MOLECULE TYPE: peptide
  - (ix) FEATURE:
    - (A) NAME/KEY: peptide
    - (B) LOCATION: 1..143
    - (D) OTHER INFORMATION: / Ceres Seq. ID 1482376
  - (xi) SEQUENCE DESCRIPTION: SEQ ID NO:517:

```
Met Arg Gly Gly Ala Met Lys Ala Arg Arg Ser Ser Thr Ser Ser
```

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1   | 5   | 10  | 15  |     |     |     |     |     |     |     |     |     |     |     |     |
| Ala | Pro | Ser | Pro | Arg | Val | Pro | Ser | Ser | Pro | Arg | Ser | Tyr | Ser | Trp | Ile |
|     | 20  |     |     |     |     |     | 25  |     |     |     |     |     | 30  |     |     |
| His | Arg | Arg | Ser | Leu | Leu | Val | Thr | Xaa | Pro | Ala | Ser | Pro | Xaa | Xaa | Ser |
|     | 35  |     |     |     |     |     | 40  |     |     |     |     | 45  |     |     |     |
| Ser | Val | Ser | Glu | Ser | Ala | Asn | Leu | Pro | Ala | Glu | Gly | Ser | Asp | Ser | Ala |
|     | 50  |     |     |     |     | 55  |     |     |     |     | 60  |     |     |     |     |
| Pro | Xaa | Ser | Val | Val | Ala | Ala | Ser | Ser | Ser | Pro | Ser | Leu | Ala | Ala | Ser |
| 65  |     |     |     |     | 70  |     |     |     |     | 75  |     |     |     |     | 80  |
| Ser | Pro | Asn | Met | Glu | Trp | Trp | Gly | Tyr | Pro | Val | Arg | Ile | Ser | Pro | Arg |
|     |     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |
| Ala | Ala | Trp | Val | Ala | Ser | Lys | Trp | Gly | Ile | Ile | Val | Gly | Leu | Leu | Asp |
|     | 100 |     |     |     |     |     |     | 105 |     |     |     |     | 110 |     |     |
| Ile | Ser | Gln | Gln | Arg | Phe | Ala | His | Leu | Phe | Pro | Lys | Leu | Cys | Cys | Ser |
|     | 115 |     |     |     |     |     | 120 |     |     |     |     | 125 |     |     |     |
| Gly | Ser | Thr | Val | Glu | Val | Val | Cys | Ser | Val | Leu | Leu | Phe | Leu |     |     |
|     | 130 |     |     |     |     | 135 |     |     |     | 140 |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:518:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 138 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..138

(D) OTHER INFORMATION: / Ceres Seq. ID 1482377

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:518:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Lys | Ala | Leu | Arg | Arg | Ser | Ser | Thr | Ser | Ser | Ala | Pro | Ser | Pro | Arg |
| 1   |     |     | 5   |     |     |     |     | 10  |     |     |     |     |     | 15  |     |
| Val | Pro | Ser | Ser | Pro | Arg | Ser | Tyr | Ser | Trp | Ile | His | Arg | Arg | Ser | Leu |
|     |     | 20  |     |     |     |     |     | 25  |     |     |     |     | 30  |     |     |
| Leu | Val | Thr | Xaa | Pro | Ala | Ser | Pro | Xaa | Xaa | Ser | Ser | Val | Ser | Glu | Ser |
|     |     | 35  |     |     |     | 40  |     |     |     |     |     | 45  |     |     |     |
| Ala | Asn | Leu | Pro | Ala | Glu | Gly | Ser | Asp | Ser | Ala | Pro | Xaa | Ser | Val | Val |
|     | 50  |     |     |     | 55  |     |     |     |     | 60  |     |     |     |     |     |
| Ala | Ala | Ser | Ser | Ser | Pro | Ser | Leu | Ala | Ala | Ser | Ser | Pro | Asn | Met | Glu |
| 65  |     |     |     |     | 70  |     |     |     |     | 75  |     |     |     |     | 80  |
| Trp | Trp | Gly | Tyr | Pro | Val | Arg | Ile | Ser | Pro | Arg | Ala | Ala | Trp | Val | Ala |
|     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |     |
| Ser | Lys | Trp | Gly | Ile | Ile | Val | Gly | Leu | Leu | Asp | Ile | Ser | Gln | Gln | Arg |
|     |     |     | 100 |     |     |     | 105 |     |     |     |     |     | 110 |     |     |
| Phe | Ala | His | Leu | Phe | Pro | Lys | Leu | Cys | Cys | Ser | Gly | Ser | Thr | Val | Glu |
|     | 115 |     |     |     |     | 120 |     |     |     |     |     | 125 |     |     |     |
| Val | Val | Val | Cys | Ser | Val | Leu | Leu | Phe | Leu |     |     |     |     |     |     |
|     | 130 |     |     |     |     | 135 |     |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:519:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 585 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..585

(D) OTHER INFORMATION: / Ceres Seq. ID 1482378

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:519:

```
aattcattac cggaagagaa aaaaataact cggaaaagaa ggagacgccg aaaattcgaa 60
aggggagggg aaagcaaagc tgatggcgga ggaccagggg aaagcaaagc aaatggcgga 120
ggccccgagc aagatcgaat ccatgaggaa gtgggtcgtc gagcacaagc tccgagccgt 180
aggttgcttc tggctaggtg ggatcagcag ttgatcgcc tacaactggg cgcgggccaa 240
tatgaagcct agcgtcaaga tcatccacgc aaggttgcat gctcaagctc taacctggc 300
tgattagtt ggttctgcat gcgtggagta ctatgacag aagtatggtt cttctgggc 360
aaaggtggac aaatacacaa gccaatacct ggcccattcg cataaagatt aaaggtcgcc 420
atgttggttc ctgcatgccg gattaatattt gggctcatct cgggttgctc atgaccgcc 480
catggatgct ggatgtttat tctttttttg tcttcataat tacaaaatgg tgggtgtactt 540
gccaggcaaa tgtaaatgag ggtataatgc agatattgtc gtcgc
```

(2) INFORMATION FOR SEQ ID NO:520:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 109 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..109

(D) OTHER INFORMATION: / Ceres Seq. ID 1482379

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:520:

```
Met Ala Glu Asp Gln Gly Lys Ala Lys Gln Met Ala Glu Ala Pro Ser
1 5 10 15
Lys Ile Glu Ser Met Arg Lys Trp Val Val Glu His Lys Leu Arg Ala
 20 25 30
Val Gly Cys Leu Trp Leu Gly Gly Ile Ser Ser Ser Ile Ala Tyr Asn
 35 40 45
Trp Ser Arg Pro Asn Met Lys Pro Ser Val Lys Ile Ile His Ala Arg
 50 55 60
Leu His Ala Gln Ala Leu Thr Leu Ala Ala Leu Val Gly Ser Ala Cys
 65 70 75 80
Val Glu Tyr Tyr Asp Gln Lys Tyr Gly Ser Ser Gly Pro Lys Val Asp
 85 90 95
Lys Tyr Thr Ser Gln Tyr Leu Ala His Ser His Lys Asp
 100 105
```

(2) INFORMATION FOR SEQ ID NO:521:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 99 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..99

(D) OTHER INFORMATION: / Ceres Seq. ID 1482380

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:521:

```
Met Ala Glu Ala Pro Ser Lys Ile Glu Ser Met Arg Lys Trp Val Val
1 5 10 15
Glu His Lys Leu Arg Ala Val Gly Cys Leu Trp Leu Gly Gly Ile Ser
 20 25 30
Ser Ser Ile Ala Tyr Asn Trp Ser Arg Pro Asn Met Lys Pro Ser Val
 35 40 45
Lys Ile Ile His Ala Arg Leu His Ala Gln Ala Leu Thr Leu Ala Ala
 50 55 60
Leu Val Gly Ser Ala Cys Val Glu Tyr Tyr Asp Gln Lys Tyr Gly Ser
 65 70 75 80
Ser Gly Pro Lys Val Asp Lys Tyr Thr Ser Gln Tyr Leu Ala His Ser
```

His Lys Asp 85 90 95

(2) INFORMATION FOR SEQ ID NO:522:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 89 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..89
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482381

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:522:

```
Met Arg Lys Trp Val Val Glu His Lys Leu Arg Ala Val Gly Cys Leu
1 5 10 15
Trp Leu Gly Gly Ile Ser Ser Ser Ile Ala Tyr Asn Trp Ser Arg Pro
20 25 30
Asn Met Lys Pro Ser Val Lys Ile Ile His Ala Arg Leu His Ala Gln
35 40 45
Ala Leu Thr Leu Ala Ala Leu Val Gly Ser Ala Cys Val Glu Tyr Tyr
50 55 60
Asp Gln Lys Tyr Gly Ser Ser Gly Pro Lys Val Asp Lys Tyr Thr Ser
65 70 75 80
Gln Tyr Leu Ala His Ser His Lys Asp
85
```

(2) INFORMATION FOR SEQ ID NO:523:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 769 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..769
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482382

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:523:

```
caaaaaaaaaac caatcgagacg gaaacgaaaa aggcctcact catctccgctc cgtccgccgc 60
accgctcgccg agcgccgctc cgcgccggag acgtcctggt tttttgccc ctacgagcgc 120
tgtccctctt ttctttccgc ggttctgccc caacttctgc atccgaatct cccacgaagt 180
tgtcacggcg atggcagcga ccggcgccgt ttcaactgac gatatcccga tcttgcaagc 240
agagaacctc accagcaacg tcaagtccgt ccactacagt cgaacattct tgtcgatcat 300
tggtggagtt gttgctggaa tctggggatt cacaggcttg acgggatttg tcttctactt 360
tctgataatg atggttgcat ctatcgggct cttagcaaa tcaaagtttt cagtgcagac 420
atacttcgat agttggacca ggatttcaat tgaaggagtt tttggtggcc ttatgtcatt 480
cgtgctgttc tggacatttg cttatgacat tgttcatatc ttctgatgga cgtagaaaga 540
gctaccctcc aaagaaaata tggaatttca tctgatgtcg aacattccca atgggctctt 600
tgtacactca gtttttattt tggtaattgt tgatataata ttttgtgata ctatatcggt 660
ggacctaaag agagctcata aactgatgta gcaactcctt cgcttgatg atctgtagca 720
gttgtgattt gtcatttcca gtaatgaatg taaactttga ttgatggac
```

(2) INFORMATION FOR SEQ ID NO:524:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 174 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..174

(D) OTHER INFORMATION: / Ceres Seq. ID 1482383

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:524:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Lys | Lys | Asn | Gln | Ser | Asp | Gly | Asn | Glu | Lys | Gly | Leu | Thr | His | Leu | Arg |
| 1   |     |     |     | 5   |     |     |     |     | 10  |     |     |     |     | 15  |     |
| Pro | Ser | Ala | Ala | Pro | Ser | Pro | Ser | Ala | Ala | Pro | Arg | Arg | Arg | Arg | Pro |
|     |     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |
| Val | Phe | Leu | Pro | Pro | Thr | Ser | Ala | Val | Pro | Leu | Phe | Phe | Pro | Arg | Phe |
|     |     | 35  |     |     |     |     | 40  |     |     |     |     | 45  |     |     |     |
| Cys | Pro | Asn | Phe | Cys | Ile | Arg | Ile | Ser | His | Glu | Val | Val | Thr | Ala | Met |
|     |     | 50  |     |     |     | 55  |     |     |     |     | 60  |     |     |     |     |
| Ala | Ala | Thr | Gly | Gly | Val | Ser | Thr | Asp | Asp | Ile | Pro | Ile | Leu | Gln | Ala |
| 65  |     |     |     |     | 70  |     |     |     |     | 75  |     |     |     |     | 80  |
| Glu | Asn | Leu | Thr | Ser | Asn | Val | Lys | Ser | Val | His | Tyr | Ser | Arg | Thr | Phe |
|     |     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |
| Leu | Ser | Ile | Ile | Gly | Gly | Val | Val | Ala | Gly | Ile | Trp | Gly | Phe | Thr | Gly |
|     |     |     | 100 |     |     |     |     | 105 |     |     |     |     | 110 |     |     |
| Leu | Thr | Gly | Phe | Val | Phe | Tyr | Phe | Leu | Ile | Met | Met | Val | Ala | Ser | Ile |
|     |     |     | 115 |     |     |     | 120 |     |     |     |     | 125 |     |     |     |
| Gly | Leu | Leu | Ala | Lys | Ser | Lys | Phe | Ser | Val | Gln | Thr | Tyr | Phe | Asp | Ser |
|     |     | 130 |     |     |     | 135 |     |     |     |     | 140 |     |     |     |     |
| Trp | Thr | Arg | Ile | Ser | Ile | Glu | Gly | Val | Phe | Gly | Gly | Leu | Met | Ser | Phe |
| 145 |     |     |     | 150 |     |     |     |     |     | 155 |     |     |     |     | 160 |
| Val | Leu | Phe | Trp | Thr | Phe | Ala | Tyr | Asp | Ile | Val | His | Ile | Phe |     |     |
|     |     |     |     | 165 |     |     |     |     | 170 |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:525:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 112 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..112

(D) OTHER INFORMATION: / Ceres Seq. ID 1482384

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:525:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Lys | Lys | Thr | Asn | Arg | Thr | Glu | Thr | Lys | Lys | Ala | Ser | Leu | Ile | Ser | Val |
| 1   |     |     |     | 5   |     |     |     |     | 10  |     |     |     |     | 15  |     |
| Arg | Pro | Pro | His | Arg | Arg | Arg | Ala | Pro | Leu | Arg | Ala | Gly | Asp | Val | Leu |
|     |     |     | 20  |     |     |     | 25  |     |     |     |     | 30  |     |     |     |
| Phe | Phe | Cys | Arg | Leu | Arg | Ala | Leu | Ser | Leu | Phe | Ser | Phe | Arg | Gly | Ser |
|     |     | 35  |     |     |     | 40  |     |     |     |     |     | 45  |     |     |     |
| Ala | Pro | Thr | Ser | Ala | Ser | Glu | Ser | Pro | Thr | Lys | Leu | Ser | Arg | Arg | Trp |
|     |     | 50  |     |     |     | 55  |     |     |     |     | 60  |     |     |     |     |
| Gln | Arg | Pro | Ala | Ala | Phe | Gln | Leu | Thr | Ile | Ser | Arg | Ser | Cys | Lys | Gln |
| 65  |     |     |     | 70  |     |     |     |     |     | 75  |     |     |     |     | 80  |
| Arg | Thr | Ser | Pro | Ala | Thr | Ser | Ser | Pro | Ser | Thr | Thr | Val | Glu | His | Ser |
|     |     |     | 85  |     |     |     |     |     | 90  |     |     |     |     | 95  |     |
| Cys | Arg | Ser | Leu | Val | Glu | Leu | Leu | Leu | Glu | Ser | Gly | Asp | Ser | Gln | Ala |
|     |     |     | 100 |     |     |     |     | 105 |     |     |     |     | 110 |     |     |

(2) INFORMATION FOR SEQ ID NO:526:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 111 amino acids

(B) TYPE: amino acid

- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
  - (A) NAME/KEY: peptide
  - (B) LOCATION: 1..111
  - (D) OTHER INFORMATION: / Ceres Seq. ID 1482385

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:526:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Ala | Ala | Thr | Gly | Gly | Val | Ser | Thr | Asp | Asp | Ile | Pro | Ile | Leu | Gln |
| 1   |     |     |     | 5   |     |     |     | 10  |     |     |     |     |     | 15  |     |
| Ala | Glu | Asn | Leu | Thr | Ser | Asn | Val | Lys | Ser | Val | His | Tyr | Ser | Arg | Thr |
|     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |     |
| Phe | Leu | Ser | Ile | Ile | Gly | Gly | Val | Ala | Gly | Ile | Trp | Gly | Phe | Thr |     |
|     | 35  |     |     |     |     | 40  |     |     |     |     | 45  |     |     |     |     |
| Gly | Leu | Thr | Gly | Phe | Val | Phe | Tyr | Phe | Leu | Ile | Met | Met | Val | Ala | Ser |
|     | 50  |     |     |     |     | 55  |     |     |     |     | 60  |     |     |     |     |
| Ile | Gly | Leu | Leu | Ala | Lys | Ser | Lys | Phe | Ser | Val | Gln | Thr | Tyr | Phe | Asp |
| 65  |     |     |     |     | 70  |     |     |     |     | 75  |     |     |     | 80  |     |
| Ser | Trp | Thr | Arg | Ile | Ser | Ile | Glu | Gly | Val | Phe | Gly | Gly | Leu | Met | Ser |
|     |     |     | 85  |     |     |     |     |     | 90  |     |     |     |     | 95  |     |
| Phe | Val | Leu | Phe | Trp | Thr | Phe | Ala | Tyr | Asp | Ile | Val | His | Ile | Phe |     |
|     |     | 100 |     |     |     |     |     | 105 |     |     |     |     |     | 110 |     |

(2) INFORMATION FOR SEQ ID NO:527:

- (i) SEQUENCE CHARACTERISTICS:
  - (A) LENGTH: 767 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- (ix) FEATURE:
  - (A) NAME/KEY: -
  - (B) LOCATION: 1..767
  - (D) OTHER INFORMATION: / Ceres Seq. ID 1482386

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:527:

|            |            |            |             |             |            |     |
|------------|------------|------------|-------------|-------------|------------|-----|
| agtcagacat | agagaatcct | tctagacaca | gcgatgtgcc  | ggtgccccag  | caattcattg | 60  |
| tggcatttcg | cacccacatc | aacccttca  | cacacgaacc  | agatcagaaa  | agccactact | 120 |
| gctttctctc | tctctctcac | acacacacag | acacaaataa  | aagaaatcag  | tagttcgatt | 180 |
| tctcctctca | cctttattta | cacatatctc | tgtattttaca | aattagggttg | ttgatgtagg | 240 |
| ctgtacgcac | ctgctagttt | gctactcgat | cctatatatc  | gtccaatcct  | atctgacctc | 300 |
| tgcacatctg | gtccttgatt | actcgtcctt | tttgcttggt  | tatatcgtcg  | ccccggccgc | 360 |
| ttgagctagc | tctctctagt | tctcgcgctc | gtcgtcgatc  | ggttgtttgc  | atagcccacg | 420 |
| gcgagccgaa | ggaataatgt | cgtcggcgcc | cctgcagatc  | gcgccggtgc  | cgggagcatg | 480 |
| tgtgctacgt | gcactgcaac | ttctgcaaca | caattctcgc  | ggtaaaccac  | ctcatctctc | 540 |
| tgtttgtccc | cctccctcct | ttgaattccc | agttctcgat  | cggcatgcat  | gcctctgaag | 600 |
| tgcagatcta | caaaggggag | atgcacatga | aatgattgst  | gcgcgcgcgc  | atgcatcata | 660 |
| cagtttattt | tgtaggattt | ggctgtcccc | tcttgctgga  | tttcttcttc  | ttcttcttta | 720 |
| tttttttgct | ctataaattg | ttttgtaaag | gttgaatgaa  | atttctg     |            |     |

(2) INFORMATION FOR SEQ ID NO:528:

- (i) SEQUENCE CHARACTERISTICS:
    - (A) LENGTH: 56 amino acids
    - (B) TYPE: amino acid
    - (C) STRANDEDNESS:
    - (D) TOPOLOGY: linear
  - (ii) MOLECULE TYPE: peptide
  - (ix) FEATURE:
    - (A) NAME/KEY: peptide
    - (B) LOCATION: 1..56
    - (D) OTHER INFORMATION: / Ceres Seq. ID 1482387
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:528:

Ser Asp Ile Glu Asn Pro Ser Arg His Ser Asp Val Pro Val Pro Gln  
1 5 10 15  
Gln Phe Ile Val Ala Phe Arg Thr His Ile Asn Pro Phe Thr His Glu  
20 25 30  
Pro Asp Gln Lys Ser His Tyr Cys Phe Leu Ser Leu Ser His Thr His  
35 40 45  
Thr Asp Thr Asn Lys Arg Asn Gln  
50 55

(2) INFORMATION FOR SEQ ID NO:529:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 66 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..66
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482388

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:529:

Met Cys Arg Cys Pro Ser Asn Ser Leu Trp His Phe Ala Pro Thr Ser  
1 5 10 15  
Thr Pro Ser His Thr Asn Gln Ile Arg Lys Ala Thr Thr Ala Phe Ser  
20 25 30  
Leu Ser Leu Thr His Thr Gln Thr Gln Ile Lys Glu Ile Ser Ser Ser  
35 40 45  
Ile Ser Pro Leu Thr Phe Ile Tyr Thr Tyr Leu Cys Ile Tyr Lys Leu  
50 55 60  
Gly Cys  
65

(2) INFORMATION FOR SEQ ID NO:530:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 53 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..53
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482389

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:530:

Met Ser Ser Ala Pro Leu Gln Ile Ala Pro Val Pro Gly Ala Cys Val  
1 5 10 15  
Leu Arg Ala Leu Gln Leu Leu Gln His Asn Ser Arg Gly Lys His Pro  
20 25 30  
His Leu Ser Val Cys Pro Pro Pro Ser Phe Glu Phe Pro Val Leu Asp  
35 40 45  
Arg His Ala Cys Leu  
50

(2) INFORMATION FOR SEQ ID NO:531:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1023 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -



(B) LOCATION: 1..1023

(D) OTHER INFORMATION: / Ceres Seq. ID 1482398

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:531:

|             |             |             |             |            |            |      |
|-------------|-------------|-------------|-------------|------------|------------|------|
| atttgccat   | ttccggtgca  | ggcagtcctg  | cagagcgagc  | aggcarrcaa | ctgggccaga | 60   |
| rtcagacag   | cctgcccgc   | acgccgtccc  | gacggccatg  | gcccgaccct | tctcttcccc | 120  |
| acacatccc   | tcttcatcct  | gggtgactcg  | ccgccctcct  | ccctggacct | cctcaccgtc | 180  |
| cgtctccacc  | acgagaacta  | gcgccctcgtc | tattaccgcg  | tgcagaccgc | gccgcagggc | 240  |
| agtcgcgcg   | gcggcctccc  | tccacctcgg  | cccgggggag  | atcgccgagc | tcgcgcgcaa | 300  |
| caaggttttr  | attgcccga   | cagtrgcgag  | cgcgatcggg  | cagctgtcca | agcccttcac | 360  |
| ctcgggtcaag | aatgggggcg  | tcggcgccgg  | ccttgacctc  | aggaccgtct | tccgctccgg | 420  |
| aggggatgccc | tccayycact  | ccgcgagtg   | tggtgcagtt  | gctacttcgc | ttgggctaga | 480  |
| aaggggggttt | rcagactcca  | tatttggaat  | gtcagtrgw   | tttkcagcaa | ttgtaatgta | 540  |
| tgatgctcag  | ggagtaagaa  | gagaaktggg  | caaccacgcc  | aagatcttga | acaggttttg | 600  |
| gacccctcaaa | gagaaggtac  | ctctggagta  | ttctgaagt   | gacatggcag | ctcctggggt | 660  |
| tgtttcggtc  | accgaggaag  | cgagctccaa  | cgcgagcccc  | tccttgaagc | gcggttctag | 720  |
| caccgaatca  | ccaaggggtga | atgggctccg  | tgggtcagag  | cctgagctga | cagagctgaa | 780  |
| gcaggcttgc  | gtagaggagg  | attaccgggt  | gagtgaatct  | gttgccaca  | cggagcttca | 840  |
| ggtcacagtc  | ggcgccctgt  | tgggttttgc  | tgtaagctta  | gcagtgtatg | caacactgta | 900  |
| acggaccttt  | tcatatcacg  | tccttgattg  | attacacatt  | tacacttttt | tttacacaga | 960  |
| aacaatacat  | gcggtttatt  | gttcccaccg  | tttaaatacag | aaatgcctat | gctagctcgt | 1020 |
| ttc         |             |             |             |            |            |      |

(2) INFORMATION FOR SEQ ID NO:532:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 299 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..299

(D) OTHER INFORMATION: / Ceres Seq. ID 1482399

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:532:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Phe | Ala | Tyr | Phe | Arg | Cys | Arg | Gln | Ser | Gly | Arg | Ala | Ser | Arg | Xaa | Xaa |
| 1   |     |     |     | 5   |     |     |     |     | 10  |     |     |     |     | 15  |     |
| Thr | Gly | Pro | Xaa | Ser | Asp | Arg | Pro | Ala | Arg | His | Ala | Val | Pro | Thr | Ala |
|     |     |     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |
| Met | Ala | Arg | Pro | Phe | Ser | Ser | Pro | His | Ile | Pro | Ser | Ser | Ser | Trp | Val |
|     |     |     |     | 35  |     |     |     |     | 40  |     |     |     |     | 45  |     |
| Thr | Arg | Arg | Pro | Pro | Pro | Trp | Thr | Ser | Ser | Pro | Ser | Val | Ser | Thr | Thr |
|     |     |     |     | 50  |     |     |     |     | 55  |     |     |     |     | 60  |     |
| Arg | Thr | Ser | Ala | Ser | Ser | Ile | Thr | Ala | Cys | Arg | Pro | Arg | Arg | Arg | Ala |
|     |     |     |     | 65  |     |     |     |     | 70  |     |     |     |     | 75  |     |
| Val | Ala | Pro | Ala | Ala | Ser | Leu | His | Leu | Gly | Pro | Gly | Glu | Ile | Ala | Glu |
|     |     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |
| Leu | Ala | Arg | Asn | Lys | Val | Xaa | Ile | Ala | Ala | Thr | Xaa | Ala | Ser | Ala | Ile |
|     |     |     |     | 100 |     |     |     |     | 105 |     |     |     |     | 110 |     |
| Gly | Gln | Leu | Ser | Lys | Pro | Phe | Thr | Ser | Val | Lys | Asn | Gly | Gly | Val | Gly |
|     |     |     |     | 115 |     |     |     |     | 120 |     |     |     |     | 125 |     |
| Ala | Gly | Leu | Asp | Leu | Arg | Thr | Val | Phe | Arg | Ser | Gly | Gly | Met | Pro | Ser |
|     |     |     |     | 130 |     |     |     |     | 135 |     |     |     |     | 140 |     |
| Xaa | His | Ser | Ala | Ser | Val | Val | Ala | Val | Ala | Thr | Ser | Leu | Gly | Leu | Glu |
|     |     |     |     | 145 |     |     |     |     | 150 |     |     |     |     | 155 |     |
| Arg | Gly | Phe | Xaa | Asp | Ser | Ile | Phe | Gly | Met | Ser | Xaa | Xaa | Phe | Xaa | Ala |
|     |     |     |     | 165 |     |     |     |     | 170 |     |     |     |     | 175 |     |
| Ile | Val | Met | Tyr | Asp | Ala | Gln | Gly | Val | Arg | Arg | Glu | Xaa | Gly | Asn | His |
|     |     |     |     | 180 |     |     |     |     | 185 |     |     |     |     | 190 |     |
| Ala | Lys | Ile | Leu | Asn | Arg | Phe | Trp | Ile | Leu | Lys | Glu | Lys | Val | Pro | Leu |
|     |     |     |     | 195 |     |     |     |     | 200 |     |     |     |     | 205 |     |

Glu Tyr Ser Glu Val Asp Met Ala Ala Pro Gly Phe Val Ser Val Thr  
210 215 220  
Glu Glu Ala Ser Ser Asn Ala Ser Pro Ser Leu Lys Arg Gly Ser Ser  
225 230 235 240  
Thr Glu Ser Pro Arg Val Asn Gly Leu Arg Gly Ser Glu Pro Glu Leu  
245 250 255  
Thr Glu Leu Lys Gln Ala Cys Val Glu Glu Asp Tyr Arg Leu Ser Glu  
260 265 270  
Ser Val Gly His Thr Glu Leu Gln Val Thr Val Gly Ala Leu Leu Gly  
275 280 285  
Phe Ala Val Ser Leu Ala Val Tyr Ala Thr Leu  
290 295

(2) INFORMATION FOR SEQ ID NO:533:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 267 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..267

(D) OTHER INFORMATION: / Ceres Seq. ID 1482400

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:533:

Met Ala Arg Pro Phe Ser Ser Pro His Ile Pro Ser Ser Ser Trp Val  
1 5 10 15  
Thr Arg Arg Pro Pro Pro Trp Thr Ser Ser Pro Ser Val Ser Thr Thr  
20 25 30  
Arg Thr Ser Ala Ser Ser Ile Thr Ala Cys Arg Pro Arg Arg Arg Ala  
35 40 45  
Val Ala Pro Ala Ala Ser Leu His Leu Gly Pro Gly Glu Ile Ala Glu  
50 55 60  
Leu Ala Arg Asn Lys Val Xaa Ile Ala Ala Thr Xaa Ala Ser Ala Ile  
65 70 75 80  
Gly Gln Leu Ser Lys Pro Phe Thr Ser Val Lys Asn Gly Gly Val Gly  
85 90 95  
Ala Gly Leu Asp Leu Arg Thr Val Phe Arg Ser Gly Gly Met Pro Ser  
100 105 110  
Xaa His Ser Ala Ser Val Val Ala Val Ala Thr Ser Leu Gly Leu Glu  
115 120 125  
Arg Gly Phe Xaa Asp Ser Ile Phe Gly Met Ser Xaa Xaa Phe Xaa Ala  
130 135 140  
Ile Val Met Tyr Asp Ala Gln Gly Val Arg Arg Glu Xaa Gly Asn His  
145 150 155 160  
Ala Lys Ile Leu Asn Arg Phe Trp Ile Leu Lys Glu Lys Val Pro Leu  
165 170 175  
Glu Tyr Ser Glu Val Asp Met Ala Ala Pro Gly Phe Val Ser Val Thr  
180 185 190  
Glu Glu Ala Ser Ser Asn Ala Ser Pro Ser Leu Lys Arg Gly Ser Ser  
195 200 205  
Thr Glu Ser Pro Arg Val Asn Gly Leu Arg Gly Ser Glu Pro Glu Leu  
210 215 220  
Thr Glu Leu Lys Gln Ala Cys Val Glu Glu Asp Tyr Arg Leu Ser Glu  
225 230 235 240  
Ser Val Gly His Thr Glu Leu Gln Val Thr Val Gly Ala Leu Leu Gly  
245 250 255  
Phe Ala Val Ser Leu Ala Val Tyr Ala Thr Leu  
260 265

(2) INFORMATION FOR SEQ ID NO:534:

- (i) SEQUENCE CHARACTERISTICS:  
    (A) LENGTH: 158 amino acids  
    (B) TYPE: amino acid  
    (C) STRANDEDNESS:  
    (D) TOPOLOGY: linear  
(ii) MOLECULE TYPE: peptide  
(ix) FEATURE:  
    (A) NAME/KEY: peptide  
    (B) LOCATION: 1..158  
    (D) OTHER INFORMATION: / Ceres Seq. ID 1482401  
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:534:

Met Pro Ser Xaa His Ser Ala Ser Val Val Ala Val Ala Thr Ser Leu  
1                  5                  10                  15  
Gly Leu Glu Arg Gly Phe Xaa Asp Ser Ile Phe Gly Met Ser Xaa Xaa  
                  20                  25                  30  
Phe Xaa Ala Ile Val Met Tyr Asp Ala Gln Gly Val Arg Arg Glu Xaa  
                  35                  40                  45  
Gly Asn His Ala Lys Ile Leu Asn Arg Phe Trp Ile Leu Lys Glu Lys  
50                  55                  60  
Val Pro Leu Glu Tyr Ser Glu Val Asp Met Ala Ala Pro Gly Phe Val  
65                  70                  75                  80  
Ser Val Thr Glu Glu Ala Ser Ser Asn Ala Ser Pro Ser Leu Lys Arg  
                  85                  90                  95  
Gly Ser Ser Thr Glu Ser Pro Arg Val Asn Gly Leu Arg Gly Ser Glu  
                  100                  105                  110  
Pro Glu Leu Thr Glu Leu Lys Gln Ala Cys Val Glu Glu Asp Tyr Arg  
                  115                  120                  125  
Leu Ser Glu Ser Val Gly His Thr Glu Leu Gln Val Thr Val Gly Ala  
130                  135                  140  
Leu Leu Gly Phe Ala Val Ser Leu Ala Val Tyr Ala Thr Leu  
145                  150                  155

- (2) INFORMATION FOR SEQ ID NO:535:  
    (i) SEQUENCE CHARACTERISTICS:  
        (A) LENGTH: 524 base pairs  
        (B) TYPE: nucleic acid  
        (C) STRANDEDNESS: single  
        (D) TOPOLOGY: linear  
    (ii) MOLECULE TYPE: DNA (genomic)  
    (ix) FEATURE:  
        (A) NAME/KEY: -  
        (B) LOCATION: 1..524  
        (D) OTHER INFORMATION: / Ceres Seq. ID 1482402  
    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:535:

ttccggctcc gctcagtcag gctcagatc ggtcgaatcc agcaccctcc ccagatttgc          60  
gtcaccaatc ttcttcttct tccgcgcgcg ccgcccgtcc cccacaagga ggtagctgc          120  
tatccccaaa tcgattcatc aatcatccgt gtccttccat ttcattccag tcggtcgccg          180  
cagcacggac cgagaacaga gcatcacgtc acatcaaact aacctaacca gctcgtccc          240  
tcgctgcgta tctgctgcac ttatcatcaac accagtcttt ctctccttg attgcattgc          300  
ccaggcaaga gaacgcacgc acaccgaccg gaatagccat gatcttctga tccaatccaa          360  
gatgggcctc aaggagcagc agctagacgc cactgaccaa actcgtgatg ccgccaactc          420  
cctcgtttct gtttctgacg agcaccacga gggaccccg gtctcaagct gcagcaccga          480  
caaggattct ggccttccaa gttgccgagt ctgccattgc gtgg

- (2) INFORMATION FOR SEQ ID NO:536:  
    (i) SEQUENCE CHARACTERISTICS:  
        (A) LENGTH: 174 amino acids  
        (B) TYPE: amino acid  
        (C) STRANDEDNESS:  
        (D) TOPOLOGY: linear  
    (ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..174

(D) OTHER INFORMATION: / Ceres Seq. ID 1482403

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:536:

```
Ser Gly Ser Ala Gln Ser Gly Leu Arg Ser Val Glu Ser Ser Thr Pro
1 5 10 15
Ser Arg Phe Ala Ser Pro Ile Phe Phe Phe Phe Arg Arg Arg Arg Arg
20 25 30
Ser Pro Thr Arg Arg Leu Ala Ala Ile Pro Lys Ser Ile His Gln Ser
35 40 45
Ser Val Ser Phe His Phe Ile Pro Val Gly Arg Arg Ser Thr Asp Arg
50 55 60
Glu Gln Ser Ile Thr Ser His Gln Thr Asn Leu Thr Ser Leu Val Pro
65 70 75 80
Arg Cys Val Ser Ala Ala Leu Ser Ser Thr Pro Val Phe Leu Leu Leu
85 90 95
Asp Cys Ile Ala Gln Ala Arg Glu Arg Thr His Thr Asp Arg Asn Ser
100 105 110
His Asp Leu Leu Ile Gln Ser Lys Met Gly Leu Lys Glu Gln Gln Leu
115 120 125
Asp Ala Thr Asp Gln Thr Arg Asp Ala Ala Asn Ser Leu Ala Ser Val
130 135 140
Ser Asp Glu His His Glu Gly Pro Arg Val Ser Ser Cys Ser Thr Asp
145 150 155 160
Lys Asp Ser Gly Leu Pro Ser Cys Arg Val Cys His Cys Val
165 170
```

(2) INFORMATION FOR SEQ ID NO:537:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 451 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..451

(D) OTHER INFORMATION: / Ceres Seq. ID 1482404

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:537:

```
gtggacgact ggcgcggtgg tgatacgctt gttactctcc caggtcgccc gtcgagtcga 60
gcttctgcgt ggagtcgctt ctgcctctgc acaccgccac cgccggtgca cgcattgacgt 120
ccatgctcgc cgctcctggc caaggcctcg gctggctcac ccaaggatct gatgaaacta 180
gatgaagagc agccggggaa agctgaaact gacggagtat aacgccccct gcgggtactt 240
taggaaccag gacatcaatt gtgcttcgag ttcttggtgt cctggaggaa acaggaagag 300
ttgattggaa aaagaaaaaa tgggatgtgt ttttcttttt gttcatgtga actgagatac 360
gacttaataa actagatctt cgaatgatgc tctgaccccc cccccccctt tttgttaatg 420
ctttttcatt gactaaaacg gttatgtaat g
```

(2) INFORMATION FOR SEQ ID NO:538:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 38 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..38

(D) OTHER INFORMATION: / Ceres Seq. ID 1482405

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:538:

Val Asp Asp Trp Arg Gly Gly Asp Thr Leu Val Thr Leu Pro Gly Arg  
1 5 10 15  
Pro Ser Ser Arg Ala Ser Ala Trp Ser Arg Phe Cys Leu Cys Thr Pro  
20 25 30  
Pro Pro Pro Val His Ala  
35

(2) INFORMATION FOR SEQ ID NO:539:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 56 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..56
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482406

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:539:

Trp Thr Thr Gly Ala Val Val Ile Arg Leu Leu Leu Ser Gln Val Ala  
1 5 10 15  
Arg Arg Val Glu Leu Leu Arg Gly Val Ala Ser Ala Ser Ala His Arg  
20 25 30  
His Arg Arg Cys Thr His Asp Val His Ala Arg Arg Ser Trp Pro Arg  
35 40 45  
Pro Arg Leu Ala His Pro Arg Ile  
50 55

(2) INFORMATION FOR SEQ ID NO:540:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 40 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..40
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482407

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:540:

Met Lys Ser Ser Arg Gly Lys Leu Lys Leu Thr Glu Tyr Asn Ala Pro  
1 5 10 15  
Cys Gly Tyr Phe Arg Asn Gln Asp Ile Asn Cys Ala Ser Ser Ser Cys  
20 25 30  
Val Pro Gly Gly Asn Arg Lys Ser  
35 40

(2) INFORMATION FOR SEQ ID NO:541:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 553 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..553
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482408

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:541:

aaagacgaag gtaatttgat cagcaccggg agacgagacg agagagagag aacgggggaa 60  
acaacaccgc cgccaggagt gttcaccggg gggaaatgat ggcgggagca ttgccaaggt 120  
ctcgccgtgc ttctccggct gcaggatggg tccgccgtgg agcatggaca gcctcgtaag 180

```
tgccttcacgc gccgggttaga aattctcttt agagttcgtc catgtattag cttcatacca 240
caccgtgtga gggggaaagg caccatcacg ccccgcggtg ngttggtcgg caatggcacc 300
gctacctgcg ctgccggcta gctatacccg aggaagaaga atagtgccag ccaatgatct 360
agaaaaagag ggcccggatt agagactagg tgaccgcttt ggctcggcca agctggaccg 420
ttgattttct gttaaaccgg catgtgtgca tacctgtgcg caatagaatg tgagccattg 480
atctgtgacg cgaggatcta cagatcatat cgtttggtga ccctggagat ctaatatgtg 540
ccatccgtgc gtg
```

(2) INFORMATION FOR SEQ ID NO:542:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 32 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..32
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482409

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:542:

```
Lys Asp Glu Gly Asn Leu Ile Ser Thr Gly Arg Arg Asp Glu Arg Glu
1 5 10 15
Arg Thr Gly Glu Thr Thr Pro Pro Pro Gly Val Phe Thr Gly Gly Lys
 20 25 30
```

(2) INFORMATION FOR SEQ ID NO:543:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 30 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..30
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482410

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:543:

```
Met Cys Ala Tyr Leu Cys Ala Ile Glu Cys Glu Pro Leu Ile Cys Asp
1 5 10 15
Pro Arg Ile Tyr Arg Ser Tyr Arg Leu Val Thr Leu Glu Ile
 20 25 30
```

(2) INFORMATION FOR SEQ ID NO:544:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 809 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..809
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482411

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:544:

```
aagtgcatta attagtgcc actgcagtag ctactagcta gcacagttca togacctcgc 60
tcgtggccgg caagcaatcg ctcaagctaa gccatggcgc ctgcgagccg cctcctcgac 120
ctggagaggc acgacgtgct cttcttctac ggcgatgggtg cctaccacca gagcgagagc 180
kncgtcgtcc ttgtcgtcgt cgtcgcgcc ctgctcctcc tcctcgtcgc gccgctcccg 240
cacgccgctg ccgtctgcgg ggcgctctac gtcgcctact gcttctcct cgaccgcgca 300
gcraagngcg agcagctcca gctcgtcgtg tccttcact gacactgccg cgcgcgcggc 360
```

```
ggcagacgcc tctcgcccca ctactcgggg cacggtggca ggctatatga tcgtgcagaa 420
gcagaattga agtcgcaatg gtcagcatgc ttatattacc agttaccatg ctttaattgca 480
tagttgcact gtagtgatca ccgcaggaag atggctctgt gtggaataga gtagtaggct 540
taagcacatt tcgtattaca ggaaaagagt ttgtggtcag aggtcttccc acgtatatag 600
ctgtctcttg agactctgca tggactctgc aatckggata tgcatgcact ataatacactt 660
cgaaataggg ccacagttga caaatcagcc aggaaacata tgtaatctgg attcttttca 720
aaaaaaaaatt gtaatacggc tactcttctc aggaatatat atgaatggac tgcacggttt 780
tctttcagtc tgttgccctgt tcttcagcc
```

(2) INFORMATION FOR SEQ ID NO:545:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 82 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..82
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482412

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:545:

```
Met Ala Pro Arg Ser Arg Leu Leu Asp Leu Glu Arg His Asp Val Leu
1 5 10 15
Phe Phe Tyr Gly Asp Gly Ala Tyr His Gln Ser Glu Ser Xaa Val Val
20 25 30
Leu Val Val Val Val Ala Ala Leu Leu Leu Leu Leu Val Ala Pro Leu
35 40 45
Pro His Ala Ala Ala Val Cys Gly Ala Leu Tyr Val Ala Tyr Cys Phe
50 55 60
Leu Leu Asp Arg Ala Xaa Lys Xaa Glu Gln Leu Gln Leu Val Val Ser
65 70 75 80
Phe His
```

(2) INFORMATION FOR SEQ ID NO:546:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 91 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..91
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482413

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:546:

```
Met Val Pro Thr Thr Arg Ala Arg Xaa Xaa Ser Ser Leu Ser Ser Ser
1 5 10 15
Ser Pro Pro Cys Ser Ser Ser Ser Ser Arg Arg Ser Arg Thr Pro Leu
20 25 30
Pro Ser Ala Gly Arg Ser Thr Ser Pro Thr Ala Ser Ser Thr Ala
35 40 45
Gln Xaa Xaa Ala Ser Ser Ser Ser Ser Ser Cys Pro Ser Thr Asp Thr
50 55 60
Ala Ala Pro Pro Ala Ala Asp Ala Ser Arg Pro Thr Thr Arg Gly Thr
65 70 75 80
Val Ala Gly Tyr Met Ile Val Gln Lys Gln Asn
85 90
```

(2) INFORMATION FOR SEQ ID NO:547:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 63 amino acids

(B) TYPE: amino acid  
(C) STRANDEDNESS:  
(D) TOPOLOGY: linear  
(ii) MOLECULE TYPE: peptide  
(ix) FEATURE:  
(A) NAME/KEY: peptide  
(B) LOCATION: 1..63  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482414  
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:547:  
Met Asp Ser Ala Ile Xaa Ile Cys Met His Tyr Asn His Phe Glu Ile  
1 5 10 15  
Gly Pro Gln Leu Thr Asn Gln Pro Gly Asn Ile Cys Asn Leu Asp Ser  
20 25 30  
Phe Gln Lys Lys Ile Val Ile Arg Leu Leu Phe Ser Gly Ile Tyr Met  
35 40 45  
Asn Gly Leu His Gly Phe Leu Ser Val Cys Cys Leu Phe Phe Ser  
50 55 60

(2) INFORMATION FOR SEQ ID NO:548:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 871 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -  
(B) LOCATION: 1..871  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482415

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:548:

atatatacgc acacgcggtg ggagtrggag ggggagactc tgccctgacc acagcaaaca 60  
acctcctctt tcctttccat ccacgcgacc atcgatcaca attttcatgg cgggtcaagga 120  
ctgcr gcg ggc cacaagg gct ggc agt ggc ggc ggc agc gr ctgtaccggc ggtgctgcgc 180  
ggcggtcgtg gctctgatcc tccgtggtcc cttcatcgtg ctgcgtcgtg ggctggtgct 240  
gcgccccac aagccccggt tctacctgca ggacctgtcg gtgctgtgcc tgaacgtgac 300  
gccgccggst ccacgtacct gttcacgacg atgcaggcga cgggtggcggc gcgcaaccgc 360  
aacgagcgcg tgggcgtgta ctacgaccag gcggacgcgt acgcggaggt acaagggcgt 420  
ggcgatcacg gtgccgacgc ggctgcccgt gcagtaccag gggccccggg acgcgtccgt 480  
gtggtccccg ttcctgcgcg ccccggaagg cggcgtgcag ytcccgcgcg agctggccgt 540  
ggcstggcgc aggacgagac ggccgggctac gtgcntstcg acgtccgcgt cgacggctgg 600  
gtccgctgga aggtcggtag cagctggatc tcgggtcact accacctccg cgtcaactgc 660  
cncgcgtgc tcaccgtcaa cgacggcagg ggcagctacg gcgccaacac cggcggcggc 720  
accggatact tccgcttcca gcaggcagsg catgcgccgt agacgtctag cagtgtcttc 780  
tctctctctg taccagctag ctgtgtttgc caattcgtcg atcgaatcaa aggacgatgc 840  
ttccttcgtc ggtgttcac actcacgcac t

(2) INFORMATION FOR SEQ ID NO:549:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 290 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS:  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide  
(B) LOCATION: 1..290  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482416

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:549:

Tyr Ile Arg Thr Arg Gly Gly Ser Xaa Arg Gly Arg Leu Cys Pro Asp  
1 5 10 15  
His Ser Lys Gln Pro Pro Leu Ser Phe Pro Ser Ile Gly Pro Ser Ile



|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |  |  |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|--|--|
|     |     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |  |  |
| Thr | Ile | Phe | Met | Ala | Val | Lys | Asp | Cys | Xaa | Gly | His | Lys | Gly | Cys | Glu |  |  |
|     |     | 35  |     |     |     |     | 40  |     |     |     |     | 45  |     |     |     |  |  |
| Cys | Glu | Arg | Glu | Xaa | Leu | Tyr | Arg | Arg | Cys | Cys | Ala | Ala | Val | Val | Ala |  |  |
|     | 50  |     |     |     |     | 55  |     |     |     |     | 60  |     |     |     |     |  |  |
| Leu | Ile | Leu | Leu | Val | Leu | Phe | Ile | Val | Leu | Val | Val | Trp | Leu | Val | Leu |  |  |
| 65  |     |     |     |     | 70  |     |     |     |     | 75  |     |     |     |     | 80  |  |  |
| Arg | Pro | His | Lys | Pro | Arg | Phe | Tyr | Leu | Gln | Asp | Leu | Ser | Val | Leu | Cys |  |  |
|     |     |     | 85  |     |     |     |     |     | 90  |     |     |     |     | 95  |     |  |  |
| Leu | Asn | Val | Thr | Pro | Pro | Xaa | Pro | Arg | Thr | Cys | Ser | Arg | Arg | Cys | Arg |  |  |
|     |     |     | 100 |     |     |     |     | 105 |     |     |     |     | 110 |     |     |  |  |
| Arg | Arg | Trp | Arg | Arg | Ala | Thr | Arg | Thr | Ser | Ala | Trp | Ala | Cys | Thr | Thr |  |  |
|     |     | 115 |     |     |     |     | 120 |     |     |     |     | 125 |     |     |     |  |  |
| Thr | Arg | Arg | Thr | Arg | Thr | Arg | Arg | Tyr | Lys | Gly | Val | Ala | Ile | Thr | Val |  |  |
|     | 130 |     |     |     |     | 135 |     |     |     |     | 140 |     |     |     |     |  |  |
| Pro | Thr | Arg | Leu | Pro | Val | Gln | Tyr | Gln | Gly | Pro | Arg | Asp | Ala | Ser | Val |  |  |
| 145 |     |     |     | 150 |     |     |     |     | 155 |     |     |     |     | 160 |     |  |  |
| Trp | Ser | Pro | Phe | Leu | Arg | Ala | Pro | Glu | Gly | Gly | Val | Gln | Xaa | Pro | Pro |  |  |
|     |     |     | 165 |     |     |     |     | 170 |     |     |     |     |     | 175 |     |  |  |
| Gln | Leu | Ala | Val | Xaa | Trp | Arg | Arg | Thr | Arg | Arg | Arg | Ala | Thr | Cys | Xaa |  |  |
|     |     |     | 180 |     |     |     |     | 185 |     |     |     |     | 190 |     |     |  |  |
| Ser | Thr | Ser | Ala | Ser | Thr | Ala | Gly | Ser | Ala | Gly | Arg | Ser | Val | Pro | Ala |  |  |
|     | 195 |     |     |     |     |     | 200 |     |     |     |     | 205 |     |     |     |  |  |
| Gly | Ser | Arg | Val | Thr | Thr | Thr | Ser | Ala | Ser | Thr | Ala | Xaa | Arg | Cys | Ser |  |  |
|     | 210 |     |     |     |     | 215 |     |     |     |     | 220 |     |     |     |     |  |  |
| Pro | Ser | Thr | Thr | Ala | Gly | Ala | Ala | Thr | Ala | Pro | Thr | Pro | Ala | Ala | Ala |  |  |
| 225 |     |     |     | 230 |     |     |     |     | 235 |     |     |     |     | 240 |     |  |  |
| Pro | Asp | Thr | Ser | Ala | Ser | Ser | Arg | Gln | Xaa | Met | Arg | Arg | Arg | Arg | Leu |  |  |
|     |     |     | 245 |     |     |     |     | 250 |     |     |     |     |     | 255 |     |  |  |
| Ala | Val | Leu | Ser | Leu | Ser | Leu | Tyr | Gln | Leu | Ala | Val | Phe | Ala | Asn | Ser |  |  |
|     |     | 260 |     |     |     |     |     | 265 |     |     |     |     | 270 |     |     |  |  |
| Ser | Ile | Glu | Ser | Lys | Asp | Asp | Ala | Ser | Phe | Val | Gly | Val | His | His | Ser |  |  |
|     | 275 |     |     |     |     |     | 280 |     |     |     |     | 285 |     |     |     |  |  |
| Arg | Thr |     |     |     |     |     |     |     |     |     |     |     |     |     |     |  |  |
|     | 290 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |  |  |

(2) INFORMATION FOR SEQ ID NO:550:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 255 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..255
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482417

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:550:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |  |  |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|--|--|
| Met | Ala | Val | Lys | Asp | Cys | Xaa | Gly | His | Lys | Gly | Cys | Glu | Cys | Glu | Arg |  |  |
| 1   |     |     | 5   |     |     |     |     | 10  |     |     |     | 15  |     |     |     |  |  |
| Glu | Xaa | Leu | Tyr | Arg | Arg | Cys | Cys | Ala | Ala | Val | Val | Ala | Leu | Ile | Leu |  |  |
|     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |     |  |  |
| Leu | Val | Leu | Phe | Ile | Val | Leu | Val | Val | Trp | Leu | Val | Leu | Arg | Pro | His |  |  |
|     | 35  |     |     |     |     | 40  |     |     |     |     | 45  |     |     |     |     |  |  |
| Lys | Pro | Arg | Phe | Tyr | Leu | Gln | Asp | Leu | Ser | Val | Leu | Cys | Leu | Asn | Val |  |  |
|     | 50  |     |     |     | 55  |     |     |     |     | 60  |     |     |     |     |     |  |  |
| Thr | Pro | Pro | Xaa | Pro | Arg | Thr | Cys | Ser | Arg | Arg | Cys | Arg | Arg | Arg | Trp |  |  |
| 65  |     |     |     | 70  |     |     |     |     | 75  |     |     |     |     | 80  |     |  |  |
| Arg | Arg | Ala | Thr | Arg | Thr | Ser | Ala | Trp | Ala | Cys | Thr | Thr | Thr | Arg | Arg |  |  |
|     |     |     | 85  |     |     |     |     | 90  |     |     |     |     |     | 95  |     |  |  |

Thr Arg Thr Arg Arg Tyr Lys Gly Val Ala Ile Thr Val Pro Thr Arg  
100 105 110  
Leu Pro Val Gln Tyr Gln Gly Pro Arg Asp Ala Ser Val Trp Ser Pro  
115 120 125  
Phe Leu Arg Ala Pro Glu Gly Gly Val Gln Xaa Pro Pro Gln Leu Ala  
130 135 140  
Val Xaa Trp Arg Arg Thr Arg Arg Arg Ala Thr Cys Xaa Ser Thr Ser  
145 150 155 160  
Ala Ser Thr Ala Gly Ser Ala Gly Arg Ser Val Pro Ala Gly Ser Arg  
165 170 175  
Val Thr Thr Thr Ser Ala Ser Thr Ala Xaa Arg Cys Ser Pro Ser Thr  
180 185 190  
Thr Ala Gly Ala Ala Thr Ala Pro Thr Pro Ala Ala Pro Asp Thr  
195 200 205  
Ser Ala Ser Ser Arg Gln Xaa Met Arg Arg Arg Arg Leu Ala Val Leu  
210 215 220  
Ser Leu Ser Leu Tyr Gln Leu Ala Val Phe Ala Asn Ser Ser Ile Glu  
225 230 235 240  
Ser Lys Asp Asp Ala Ser Phe Val Gly Val His His Ser Arg Thr  
245 250 255

(2) INFORMATION FOR SEQ ID NO:551:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 725 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..725
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482418

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:551:

|                                                                    |     |
|--------------------------------------------------------------------|-----|
| aggaacttgt aacctggctc gcagcggtgc gtgaaggacc tcgcgcgcgc tctcctctac  | 60  |
| tgcttggtcg tctcggtgcc ccggccgaac atccaagcct ctccatgtct ggcccttcga  | 120 |
| aggagcagcg cgnatgccg gcaactgggt gctggctaata ggctgtcggc acctccgct   | 180 |
| tggccttcac ctggctcgtc ttcttcggct ccgggtygct ctgctcagcc acctactccg  | 240 |
| agatacaggt gatcgccgtg catgggcgca cggttgcggg gtggacgctg ctgtcgtgca  | 300 |
| ccctctgctt cctgtgcgcc ttcaacctca ccagcanagc cgctgtacgc ggccaccttc  | 360 |
| ctgtccttcg tctacgcctt cgggtacctg agcaccgagt gcatgggtgta ccacaccatg | 420 |
| agtgcagcta gtctcgtccc gttcaccttc atcgtgtgta catccatggt ctggatgctg  | 480 |
| attcaatgga actcggatgg tcacggcccc cgtcttcttc atgggtctac tgcttccaag  | 540 |
| cagccatgac ttgcgaggtt ctctcaccta tggcttccct caactacata cggttcagtg  | 600 |
| catgcaagca ccatggaatt atggaatatc tgtaatcttt tgtaataatc gtttctatgt  | 660 |
| cccgaggct agtgaatgaa actagcaagc tatcatctgt gataaatttg taattttacc   | 720 |
| actct                                                              |     |

(2) INFORMATION FOR SEQ ID NO:552:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 142 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..142
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482419

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:552:

Glu Leu Val Thr Trp Leu Ala Ala Leu Arg Glu Gly Pro Arg Ala Arg  
1 5 10 15

Ser Pro Leu Leu Leu Gly Arg Leu Val Ala Pro Ala Glu His Pro Ser  
20 25 30  
Leu Ser Met Ser Gly Pro Ser Lys Glu Gln Arg Xaa Met Pro Ala Leu  
35 40 45  
Gly Cys Trp Leu Met Ala Val Gly Thr Phe Arg Leu Ala Phe Thr Trp  
50 55 60  
Ser Cys Phe Phe Gly Ser Gly Xaa Leu Cys Ser Ala Thr Tyr Ser Glu  
65 70 75 80  
Ile Gln Val Ile Gly Val His Gly Arg Thr Val Ala Val Trp Thr Leu  
85 90 95  
Leu Ser Cys Thr Leu Cys Phe Leu Cys Ala Phe Asn Leu Thr Ser Xaa  
100 105 110  
Ala Ala Val Arg Gly His Leu Pro Val Leu Arg Leu Arg Leu Arg Val  
115 120 125  
Pro Glu His Arg Val His Gly Val Pro His His Glu Cys Ser  
130 135 140

(2) INFORMATION FOR SEQ ID NO:553:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 108 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..108

(D) OTHER INFORMATION: / Ceres Seq. ID 1482420

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:553:

Met Ser Gly Pro Ser Lys Glu Gln Arg Xaa Met Pro Ala Leu Gly Cys  
1 5 10 15  
Trp Leu Met Ala Val Gly Thr Phe Arg Leu Ala Phe Thr Trp Ser Cys  
20 25 30  
Phe Phe Gly Ser Gly Xaa Leu Cys Ser Ala Thr Tyr Ser Glu Ile Gln  
35 40 45  
Val Ile Gly Val His Gly Arg Thr Val Ala Val Trp Thr Leu Leu Ser  
50 55 60  
Cys Thr Leu Cys Phe Leu Cys Ala Phe Asn Leu Thr Ser Xaa Ala Ala  
65 70 75 80  
Val Arg Gly His Leu Pro Val Leu Arg Leu Arg Leu Arg Val Pro Glu  
85 90 95  
His Arg Val His Gly Val Pro His His Glu Cys Ser  
100 105

(2) INFORMATION FOR SEQ ID NO:554:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 98 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..98

(D) OTHER INFORMATION: / Ceres Seq. ID 1482421

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:554:

Met Pro Ala Leu Gly Cys Trp Leu Met Ala Val Gly Thr Phe Arg Leu  
1 5 10 15  
Ala Phe Thr Trp Ser Cys Phe Phe Gly Ser Gly Xaa Leu Cys Ser Ala  
20 25 30  
Thr Tyr Ser Glu Ile Gln Val Ile Gly Val His Gly Arg Thr Val Ala

[illegible]

(2) INFORMATION FOR SEO ID NO:555:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 119 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -  
(B) LOCATION: 1..119  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482422

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:555:

aaccgcaagc tcaagcaaaa acacaaagcg cttaaaccac actcaaacca accgccagcc 60  
aacaacacag cctcctaagc ccgaccagaa ctcgctcgta gcccccagaa cccgacagc

(2) INFORMATION FOR SEQ ID NO:556:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 39 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS:  
(D) TOPOLOGY: linear

## (ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide  
(B) LOCATION: 1..39  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482423

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:556:

[illegible]

(2) INFORMATION FOR SEO ID NO:557:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 297 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -  
(B) LOCATION: 1..297  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482424

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:557:

|             |            |             |             |            |             |     |
|-------------|------------|-------------|-------------|------------|-------------|-----|
| atcttttcgcc | cgcgcgcccc | agtcccgcgtc | cgaagctgtg  | cctcgtacca | tttcgatcca  | 60  |
| atggcgccga  | cgtcgaagct | gtcgcacgggc | atcaagcgcg  | cttcgcggtc | gcacgcgtac  | 120 |
| catcgccgtg  | ggctgtgggc | catgatgaac  | ttgagcgcaa  | gaagagtata | cgcgttagtag | 180 |
| ttactctgtg  | acgtacgcag | gcagagagcg  | cgcgtctccag | cgtatacgtg | cacctgagacg | 240 |
| tagtacgtac  | atgtactacc | cgttactttgc | tctccaatcg  | agtgcacgtt | qcaqccc     |     |

(2) INFORMATION FOR SEQ ID NO:558:

(i) SEQUENCE CHARACTERISTICS:  
    (A) LENGTH: 59 amino acids  
    (B) TYPE: amino acid  
    (C) STRANDEDNESS:  
    (D) TOPOLOGY: linear  
(ii) MOLECULE TYPE: peptide  
(ix) FEATURE:  
    (A) NAME/KEY: peptide  
    (B) LOCATION: 1..59  
    (D) OTHER INFORMATION: / Ceres Seq. ID 1482425  
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:558:  
Ile Phe Arg Pro Pro Pro Val Pro Ile Arg Ser Cys Ala Ser Tyr  
1                  5                  10                  15  
His Phe Asp Pro Met Ala Pro Thr Ser Lys Leu Ser Thr Gly Ile Lys  
                  20                  25                  30  
Arg Ala Ser Arg Ser His Ala Tyr His Arg Arg Gly Leu Trp Ala Met  
                  35                  40                  45  
Met Asn Leu Ser Ala Arg Arg Val Tyr Arg Leu  
50                  55  
(2) INFORMATION FOR SEQ ID NO:559:  
    (i) SEQUENCE CHARACTERISTICS:  
        (A) LENGTH: 47 amino acids  
        (B) TYPE: amino acid  
        (C) STRANDEDNESS:  
        (D) TOPOLOGY: linear  
    (ii) MOLECULE TYPE: peptide  
    (ix) FEATURE:  
        (A) NAME/KEY: peptide  
        (B) LOCATION: 1..47  
        (D) OTHER INFORMATION: / Ceres Seq. ID 1482426  
    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:559:  
Phe Phe Ala Arg Arg Pro Gln Ser Arg Ser Glu Ala Val Pro Arg Thr  
1                  5                  10                  15  
Ile Ser Ile Gln Trp Arg Arg Arg Arg Ser Cys Arg Arg Ala Ser Ser  
                  20                  25                  30  
Ala Leu Arg Gly Arg Thr Arg Thr Ile Ala Val Gly Cys Gly Pro  
                  35                  40                  45  
(2) INFORMATION FOR SEQ ID NO:560:  
    (i) SEQUENCE CHARACTERISTICS:  
        (A) LENGTH: 62 amino acids  
        (B) TYPE: amino acid  
        (C) STRANDEDNESS:  
        (D) TOPOLOGY: linear  
    (ii) MOLECULE TYPE: peptide  
    (ix) FEATURE:  
        (A) NAME/KEY: peptide  
        (B) LOCATION: 1..62  
        (D) OTHER INFORMATION: / Ceres Seq. ID 1482427  
    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:560:  
Phe Ser Pro Ala Ala Pro Ser Pro Asp Pro Lys Leu Cys Leu Val Pro  
1                  5                  10                  15  
Phe Arg Ser Asn Gly Ala Asp Val Glu Ala Val Asp Gly His Gln Ala  
                  20                  25                  30  
Arg Phe Ala Val Ala Arg Val Pro Ser Pro Trp Ala Val Gly His Asp  
                  35                  40                  45  
Glu Leu Glu Arg Lys Lys Ser Ile Pro Leu Val Val Thr Leu  
50                  55                  60  
(2) INFORMATION FOR SEQ ID NO:561:  
    (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 606 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -

(B) LOCATION: 1..606

(D) OTHER INFORMATION: / Ceres Seq. ID 1482428

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:561:

|            |            |            |            |            |            |     |
|------------|------------|------------|------------|------------|------------|-----|
| gtctcacaaa | ctttttttaa | gtctatcggt | aaccgcttc  | agctagcgag | cattgagcag | 60  |
| tctgcagtcg | ccgagccgcg | tgtccgccgg | gccggcggtt | accaagctca | ccaaaaatct | 120 |
| ttccaggttc | gaggccgccc | atgttcgcct | gcaggtctct | cctcgcaagg | gatcatattg | 180 |
| tcgaaataag | ttggcctcat | tcgtgatgga | aggggcgcaa | ggatcaagca | ttgtgacaaa | 240 |
| acacaataaa | aggcagtcct | ctgtgcagag | atggaggcca | gtttcaacag | aagcagttcc | 300 |
| ccagcatcac | caagatgaca | ttattgagac | atcaaattct | ggaagcaaga | aaattataga | 360 |
| ggattgcata | gcttctagt  | agaatttgcc | accagatgga | acaaccaatg | ttgttgaagt | 420 |
| taccgccaat | gatgcttcat | cgtcaaaaaa | taatttaagt | tttgggtaca | gttcaactaa | 480 |
| agtagttata | gaagaccatg | cggagttatc | tggcttcaat | aaggatctag | ctgggtccaa | 540 |
| tgtcttcggg | acacattcct | ycctctgttg | ggcggktcaa | agtcgacagc | ttgactactc | 600 |
| tcattt     |            |            |            |            |            |     |

(2) INFORMATION FOR SEQ ID NO:562:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 133 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..133

(D) OTHER INFORMATION: / Ceres Seq. ID 1482429

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:562:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Glu | Gly | Ala | Gln | Gly | Ser | Ser | Ile | Val | Thr | Lys | His | Asn | Lys | Arg |
| 1   |     |     |     | 5   |     |     |     |     | 10  |     |     |     |     | 15  |     |
| Gln | Ser | Pro | Val | Gln | Arg | Trp | Arg | Pro | Val | Ser | Thr | Glu | Ala | Val | Pro |
|     |     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |
| Gln | His | His | Gln | Asp | Asp | Ile | Ile | Glu | Thr | Ser | Asn | Ser | Gly | Ser | Lys |
|     |     |     | 35  |     |     |     | 40  |     |     |     |     | 45  |     |     |     |
| Lys | Ile | Ile | Glu | Asp | Cys | Ile | Ala | Ser | Ser | Glu | Asn | Leu | Pro | Pro | Asp |
|     | 50  |     |     |     |     | 55  |     |     |     | 60  |     |     |     |     |     |
| Gly | Thr | Thr | Asn | Val | Val | Glu | Val | Thr | Ala | Asn | Asp | Ala | Ser | Ser | Ser |
| 65  |     |     |     | 70  |     |     |     | 75  |     |     |     |     | 80  |     |     |
| Lys | Asn | Asn | Leu | Ser | Phe | Gly | Tyr | Ser | Ser | Thr | Lys | Val | Val | Ile | Glu |
|     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |     |
| Asp | His | Ala | Glu | Leu | Ser | Gly | Phe | Asn | Lys | Asp | Leu | Ala | Gly | Ser | Asn |
|     |     | 100 |     |     |     |     | 105 |     |     |     |     | 110 |     |     |     |
| Val | Phe | Gly | Thr | His | Ser | Xaa | Ser | Val | Glu | Ala | Xaa | Gln | Ser | Arg | Gln |
|     |     | 115 |     |     |     | 120 |     |     |     |     |     | 125 |     |     |     |
| Leu | Asp | Tyr | Ser | His |     |     |     |     |     |     |     |     |     |     |     |
|     | 130 |     |     |     |     |     |     |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:563:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 451 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -  
(B) LOCATION: 1..451  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482430

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:563:

|            |            |            |            |            |            |     |
|------------|------------|------------|------------|------------|------------|-----|
| gaaaacggcg | atgttggccg | tcccattttg | taagctcccc | ttccccgtct | ggccgtctcg | 60  |
| actgccccag | tcctttctca | gatccatgtc | taccagaat  | ctgctaactg | gcgcgtgcac | 120 |
| gagctccgcc | ccgaccccg  | ccgaggcgga | ggaaggggac | aggacgcctt | tggtgcacgc | 180 |
| tgcgaacgcg | gcggaagagc | tgtaccgcct | ccgtgacacc | tttttcccg  | gggacccttc | 240 |
| cgagaaagtc | gccgcactcc | gcgcccgcgc | cgacgccgcc | ctcgcgctcc | tcgacgcctt | 300 |
| cccgtccgaa | caaaagaagt | ctcgacaact | gcgtggtgtt | tatgaatttt | tgaggggaaa | 360 |
| aatactggat | gtctttcctg | attatcataa | ggaggctgaa | gatcatttat | ccaaagcagt | 420 |
| aaagttgaac | ccatctcttg | tagatgcatg | g          |            |            |     |

(2) INFORMATION FOR SEQ ID NO:564:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 150 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS:  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide  
(B) LOCATION: 1..150  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482431

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:564:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Lys | Thr | Ala | Met | Leu | Ala | Val | Pro | Phe | Cys | Lys | Leu | Pro | Phe | Pro | Val |
| 1   |     |     | 5   |     |     |     |     |     | 10  |     |     |     |     | 15  |     |
| Trp | Pro | Ser | Arg | Leu | Pro | Gln | Ser | Phe | Leu | Arg | Ser | Met | Ser | Thr | Gln |
|     |     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |
| Asn | Leu | Leu | Thr | Gly | Ala | Cys | Thr | Ser | Ser | Ala | Pro | Thr | Pro | Ser | Glu |
|     |     |     | 35  |     |     |     | 40  |     |     |     |     | 45  |     |     |     |
| Ala | Glu | Gly | Asp | Arg | Thr | Pro | Leu | Ala | Asp | Ala | Ala | Asn | Ala | Ala |     |
|     | 50  |     |     |     |     | 55  |     |     |     | 60  |     |     |     |     |     |
| Glu | Glu | Leu | Tyr | Arg | Leu | Arg | Asp | Thr | Phe | Phe | Pro | Arg | Asp | Pro | Ser |
| 65  |     |     |     | 70  |     |     |     |     | 75  |     |     |     |     | 80  |     |
| Glu | Lys | Val | Ala | Ala | Leu | Arg | Ala | Arg | Ala | Asp | Ala | Ala | Leu | Ala | Leu |
|     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |     |
| Leu | Asp | Ala | Phe | Pro | Ser | Glu | Gln | Lys | Lys | Ser | Arg | Gln | Leu | Arg | Gly |
|     |     |     | 100 |     |     |     |     | 105 |     |     |     |     | 110 |     |     |
| Val | Tyr | Glu | Phe | Leu | Arg | Gly | Lys | Ile | Leu | Asp | Val | Phe | Pro | Asp | Tyr |
|     |     | 115 |     |     |     | 120 |     |     |     |     | 125 |     |     |     |     |
| His | Lys | Glu | Ala | Glu | Asp | His | Leu | Ser | Lys | Ala | Val | Lys | Leu | Asn | Pro |
|     | 130 |     |     |     |     | 135 |     |     |     |     | 140 |     |     |     |     |
| Ser | Leu | Val | Asp | Ala | Trp |     |     |     |     |     |     |     |     |     |     |
| 145 |     |     |     |     | 150 |     |     |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:565:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 147 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS:  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide  
(B) LOCATION: 1..147  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482432

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:565:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Leu | Ala | Val | Pro | Phe | Cys | Lys | Leu | Pro | Phe | Pro | Val | Trp | Pro | Ser |
| 1   |     |     | 5   |     |     |     |     | 10  |     |     |     |     | 15  |     |     |
| Arg | Leu | Pro | Gln | Ser | Phe | Leu | Arg | Ser | Met | Ser | Thr | Gln | Asn | Leu | Leu |

|                                                                 |     |    |     |    |     |
|-----------------------------------------------------------------|-----|----|-----|----|-----|
|                                                                 | 20  |    | 25  |    | 30  |
| Thr Gly Ala Cys Thr Ser Ser Ala Pro Thr Pro Ser Glu Ala Glu Glu |     |    |     |    |     |
|                                                                 | 35  |    | 40  |    | 45  |
| Gly Asp Arg Thr Pro Leu Ala Asp Ala Ala Asn Ala Ala Glu Glu Leu |     |    |     |    |     |
|                                                                 | 50  |    | 55  |    | 60  |
| Tyr Arg Leu Arg Asp Thr Phe Phe Pro Arg Asp Pro Ser Glu Lys Val |     |    |     |    |     |
|                                                                 | 65  |    | 70  |    | 75  |
| Ala Ala Leu Arg Ala Arg Ala Asp Ala Ala Leu Ala Leu Leu Asp Ala |     |    |     |    |     |
|                                                                 |     | 85 |     | 90 |     |
| Phe Pro Ser Glu Gln Lys Lys Ser Arg Gln Leu Arg Gly Val Tyr Glu |     |    |     |    |     |
|                                                                 | 100 |    | 105 |    | 110 |
| Phe Leu Arg Gly Lys Ile Leu Asp Val Phe Pro Asp Tyr His Lys Glu |     |    |     |    |     |
|                                                                 | 115 |    | 120 |    | 125 |
| Ala Glu Asp His Leu Ser Lys Ala Val Lys Leu Asn Pro Ser Leu Val |     |    |     |    |     |
|                                                                 | 130 |    | 135 |    | 140 |
| Asp Ala Trp                                                     |     |    |     |    |     |
| 145                                                             |     |    |     |    |     |

(2) INFORMATION FOR SEQ ID NO:566:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 122 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..122
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482433

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:566:

|                                                                 |     |   |     |    |     |
|-----------------------------------------------------------------|-----|---|-----|----|-----|
| Met Ser Thr Gln Asn Leu Leu Thr Gly Ala Cys Thr Ser Ser Ala Pro |     |   |     |    |     |
| 1                                                               |     | 5 |     | 10 | 15  |
| Thr Pro Ser Glu Ala Glu Glu Gly Asp Arg Thr Pro Leu Ala Asp Ala |     |   |     |    |     |
|                                                                 | 20  |   | 25  |    | 30  |
| Ala Asn Ala Ala Glu Glu Leu Tyr Arg Leu Arg Asp Thr Phe Phe Pro |     |   |     |    |     |
|                                                                 | 35  |   | 40  |    | 45  |
| Arg Asp Pro Ser Glu Lys Val Ala Ala Leu Arg Ala Arg Ala Asp Ala |     |   |     |    |     |
|                                                                 | 50  |   | 55  |    | 60  |
| Ala Leu Ala Leu Leu Asp Ala Phe Pro Ser Glu Gln Lys Lys Ser Arg |     |   |     |    |     |
|                                                                 | 65  |   | 70  |    | 75  |
| Gln Leu Arg Gly Val Tyr Glu Phe Leu Arg Gly Lys Ile Leu Asp Val |     |   |     |    |     |
|                                                                 | 85  |   | 90  |    | 95  |
| Phe Pro Asp Tyr His Lys Glu Ala Glu Asp His Leu Ser Lys Ala Val |     |   |     |    |     |
|                                                                 | 100 |   | 105 |    | 110 |
| Lys Leu Asn Pro Ser Leu Val Asp Ala Trp                         |     |   |     |    |     |
|                                                                 | 115 |   | 120 |    |     |

(2) INFORMATION FOR SEQ ID NO:567:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 463 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..463
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482434

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:567:

|                                                                   |     |
|-------------------------------------------------------------------|-----|
| ataccgggat gggcgccatg ggcacgtcg tcagagcgtg ggcgcctccc gctcttgccg  | 60  |
| ccgcggccgc gccggcaccc tccggctcca gggacaccgg tcaggcccag cggaggagca | 120 |



```
agccctcgag gaccggccgc gtgcgcgtgc tcggcggcac tggccgtgtc ggaggatcca 180
cggccaccgc actctccaaa ctccgccccca agcttggcat cctcgtcggg ggcaggaacc 240
gggagaaagg cgagtccatt gcagccaagc ttggggggcca gtctgagttc gtccagggtcg 300
acacccgcaa cacaggcatg ttggaggaag cgctgcaggt ggtagctgtt cgcgagattg 360
ccaaaccgga ggcagctgcg acgcccggcga ggcgctcgcg ccccatccct ctggcttccg 420
tggccgtgtg gagtctggtt gccactggcg ccgcaaagtc tgc
```

(2) INFORMATION FOR SEQ ID NO:568:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 153 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..153
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482435

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:568:

```
Thr Gly Met Gly Ala Met Gly Ile Val Val Arg Ala Trp Ala Pro Pro
1 5 10 15
Ala Leu Ala Ala Ala Ala Ala Pro Ala Pro Ser Gly Ser Arg Asp Thr
 20 25 30
Gly Gln Ala Gln Arg Arg Ser Lys Pro Ser Arg Thr Gly Arg Val Arg
 35 40 45
Val Leu Gly Gly Thr Gly Arg Val Gly Gly Ser Thr Ala Thr Ala Leu
 50 55 60
Ser Lys Leu Arg Pro Lys Leu Gly Ile Leu Val Gly Gly Arg Asn Arg
 65 70 75 80
Glu Lys Gly Glu Ser Ile Ala Ala Lys Leu Gly Gly Gln Ser Glu Phe
 85 90 95
Val Gln Val Asp Thr Arg Asn Thr Gly Met Leu Glu Glu Ala Leu Gln
 100 105 110
Val Val Ala Val Arg Gly Val Ala Lys Pro Glu Ala Ala Thr Pro
 115 120 125
Ala Arg Arg Ser Arg Pro Ile Pro Leu Ala Ser Val Ala Val Trp Ser
 130 135 140
Leu Val Ala Thr Gly Ala Ala Asn Ala
 145 150
```

(2) INFORMATION FOR SEQ ID NO:569:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 151 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..151
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482436

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:569:

```
Met Gly Ala Met Gly Ile Val Val Arg Ala Trp Ala Pro Pro Ala Leu
1 5 10 15
Ala Ala Ala Ala Ala Pro Ala Pro Ser Gly Ser Arg Asp Thr Gly Gln
 20 25 30
Ala Gln Arg Arg Ser Lys Pro Ser Arg Thr Gly Arg Val Arg Val Leu
 35 40 45
Gly Gly Thr Gly Arg Val Gly Gly Ser Thr Ala Thr Ala Leu Ser Lys
 50 55 60
Leu Arg Pro Lys Leu Gly Ile Leu Val Gly Gly Arg Asn Arg Glu Lys
```

(2) INFORMATION FOR SEQ ID NO:570:

(A) LENGTH: 148 amino acids

(C) STRANDEDNESS:

(D) TOPOLOGY: lin

MOLECULE TYPE: pentide

(A) NAM

(B) LOCATION: 1..148

(D) OTHER INFORMATION

SEQUENCE DESCRIPTION: SEO ID NO:570:

Gly Ile Val Val Arg Ala Trp Ala Pro Pro A

(2) INFORMATION FOR SEQ ID NO:571:

(A) LENGTH: 511 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

MOLECULE TYPE: DNA (α)

(ix) FEATURE:

(A) NAME

(B) LOCATION: 1

(D) OTHER INFORMATION: \_\_\_\_\_

SEQUENCE DESCRIPTION: SEQ ID NO:571:

(X1) SEQUENCE DESCRIPTION: SEQ ID NO:371:  
cgacgcc gaggggtttct aacaacgtaa aaagaagaag g

|            |            |             |            |            |            |     |
|------------|------------|-------------|------------|------------|------------|-----|
| aaacgacgcc | gaggggttct | aaacacgctaa | aaagaaagaa | gaacagacag | catcaggccc | 180 |
| cgcccgtagc | tacaggggaa | aggacaaaag  | gcttcggggc | gtggcgggcc | cgctggtcga | 120 |
| cgatcgttca | gagcgcgggg | agggagaaga  | ggtcgymgyc | kscsatgtmt | sykrarcsgc | 180 |
| agccgctgac | tcctggccac | atcgggaggc  | tgccgaggtc | gagcgcgggg | tcsgcggsa  | 240 |

ggtagtcgag gccgagctgc ggmgtgggca agtcgtcgtc gaacgggacg ccgccgtaaa 300  
gagaacgcgt cctcgccgag ctggggcagg agcgcgtcat cggcggaraa cgggkwwkagg 360  
ccgccgsggc cgtcggggcg ktcctttcttg gtcgcacaac cggmggcggm gtycgyykat 420  
tttyggcggc agartckcac gcgccgtctc gtcgggcatt gcccgaggga cggagaccgg 480  
cgagccnacc accakctggg actcgtcgca g

(2) INFORMATION FOR SEQ ID NO:572:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 80 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..80

(D) OTHER INFORMATION: / Ceres Seq. ID 1482439

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:572:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Thr | Thr | Pro | Arg | Val | Ser | Asn | Asn | Val | Lys | Arg | Arg | Arg | Lys | Glu | Gln |
| 1   |     |     |     | 5   |     |     |     | 10  |     |     |     |     | 15  |     |     |
| His | Leu | Ala | Pro | Ser | Val | Arg | Thr | Gly | Glu | Arg | Thr | Lys | Gly | Phe | Gly |
|     |     | 20  |     |     |     |     |     | 25  |     |     |     | 30  |     |     |     |
| Arg | Trp | Arg | Pro | Arg | Trp | Ser | Thr | Ile | Val | Gln | Ser | Ala | Gly | Arg | Glu |
|     | 35  |     |     |     |     | 40  |     |     |     |     | 45  |     |     |     |     |
| Lys | Arg | Ser | Xaa | Xaa | Xaa | Met | Xaa | Xaa | Xaa | Xaa | Gln | Pro | Ser | Thr | Pro |
|     | 50  |     |     |     |     | 55  |     |     |     |     | 60  |     |     |     |     |
| Gly | His | Ile | Gly | Arg | Leu | Pro | Arg | Ser | Ser | Ala | Gly | Xaa | Ala | Xaa | Arg |
| 65  |     |     |     |     | 70  |     |     |     |     | 75  |     |     |     |     | 80  |

(2) INFORMATION FOR SEQ ID NO:573:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 613 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -

(B) LOCATION: 1..613

(D) OTHER INFORMATION: / Ceres Seq. ID 1482444

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:573:

|            |             |             |            |            |              |     |
|------------|-------------|-------------|------------|------------|--------------|-----|
| mrgtsccaag | aatgtttctca | cgaagctgat  | taaatcattg | aaccttagat | taactgctgt   | 60  |
| gcaactaatt | gattcccatt  | tatgtttgtga | tcccggaac  | tacgtaagtt | cgctacttct   | 120 |
| ctccttatcc | acaatgcttc  | acatggaact  | cccacatgct | aatatatatt | gtctaaaatcga | 180 |
| tctgattgga | agctacggga  | agctagcttt  | caatttagat | ttctataacc | gtagttcaaga  | 240 |
| cttgatcatc | ttggagcacc  | atcttagtca  | agatcctcgc | tctgctaagt | acagaaaact   | 300 |
| aacaaaagag | ctatgtagt   | tcattgaaga  | ttacagtctt | gttaatttta | caaccttgga   | 360 |
| tattcaggat | aaagaaagt   | ttggggatct  | agtaaagctc | atcgacaaga | gcaatggata   | 420 |
| catatttgcc | ggcattgat   | caagtgtggt  | tgaatacagc | aagattgcaa | ttgggtcaaac  | 480 |
| tgattgggat | tataacagag  | tcgcagctgt  | acaggagaag | tacatggaag | atgaggaaat   | 540 |
| acaagactga | gaacagtgt   | tgaactttta  | tatagaagag | agctggtcta | aaatatctct   | 600 |
| gaaccaaacc | att         |             |            |            |              |     |

(2) INFORMATION FOR SEQ ID NO:574:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 182 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..182

(D) OTHER INFORMATION: / Ceres Seq. ID 1482445

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:574:

Xaa Xaa Lys Asn Val Leu Thr Lys Leu Ile Lys Ser Leu Asn Leu Arg  
1 5 10 15  
Leu Thr Ala Val Gln Leu Ile Asp Ser His Leu Cys Cys Asp Pro Gly  
20 25 30  
Asn Tyr Val Ser Ser Leu Leu Leu Ser Leu Ser Thr Met Leu His Met  
35 40 45  
Glu Leu Pro His Val Asn Ile Leu Ser Lys Ile Asp Leu Ile Gly Ser  
50 55 60  
Tyr Gly Lys Leu Ala Phe Asn Leu Asp Phe Tyr Thr Asp Val Gln Asp  
65 70 75 80  
Leu Ser Tyr Leu Glu His His Leu Ser Gln Asp Pro Arg Ser Ala Lys  
85 90 95  
Tyr Arg Lys Leu Thr Lys Glu Leu Cys Ser Val Ile Glu Asp Tyr Ser  
100 105 110  
Leu Val Asn Phe Thr Thr Leu Asp Ile Gln Asp Lys Glu Ser Val Gly  
115 120 125  
Asp Leu Val Lys Leu Ile Asp Lys Ser Asn Gly Tyr Ile Phe Ala Gly  
130 135 140  
Ile Asp Ala Ser Val Val Glu Tyr Ser Lys Ile Ala Ile Gly Gln Thr  
145 150 155 160  
Asp Trp Asp Tyr Asn Arg Val Ala Ala Val Gln Glu Lys Tyr Met Glu  
165 170 175  
Asp Glu Glu Ile Gln Asp  
180

(2) INFORMATION FOR SEQ ID NO:575:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 138 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..138

(D) OTHER INFORMATION: / Ceres Seq. ID 1482446

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:575:

Met Leu His Met Glu Leu Pro His Val Asn Ile Leu Ser Lys Ile Asp  
1 5 10 15  
Leu Ile Gly Ser Tyr Gly Lys Leu Ala Phe Asn Leu Asp Phe Tyr Thr  
20 25 30  
Asp Val Gln Asp Leu Ser Tyr Leu Glu His His Leu Ser Gln Asp Pro  
35 40 45  
Arg Ser Ala Lys Tyr Arg Lys Leu Thr Lys Glu Leu Cys Ser Val Ile  
50 55 60  
Glu Asp Tyr Ser Leu Val Asn Phe Thr Thr Leu Asp Ile Gln Asp Lys  
65 70 75 80  
Glu Ser Val Gly Asp Leu Val Lys Leu Ile Asp Lys Ser Asn Gly Tyr  
85 90 95  
Ile Phe Ala Gly Ile Asp Ala Ser Val Val Glu Tyr Ser Lys Ile Ala  
100 105 110  
Ile Gly Gln Thr Asp Trp Asp Tyr Asn Arg Val Ala Ala Val Gln Glu  
115 120 125  
Lys Tyr Met Glu Asp Glu Glu Ile Gln Asp  
130 135

(2) INFORMATION FOR SEQ ID NO:576:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 135 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..135
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482447

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:576:

Met Glu Leu Pro His Val Asn Ile Leu Ser Lys Ile Asp Leu Ile Gly  
1 5 10 15  
Ser Tyr Gly Lys Leu Ala Phe Asn Leu Asp Phe Tyr Thr Asp Val Gln  
20 25 30  
Asp Leu Ser Tyr Leu Glu His His Leu Ser Gln Asp Pro Arg Ser Ala  
35 40 45  
Lys Tyr Arg Lys Leu Thr Lys Glu Leu Cys Ser Val Ile Glu Asp Tyr  
50 55 60  
Ser Leu Val Asn Phe Thr Thr Leu Asp Ile Gln Asp Lys Glu Ser Val  
65 70 75 80  
Gly Asp Leu Val Lys Leu Ile Asp Lys Ser Asn Gly Tyr Ile Phe Ala  
85 90 95  
Gly Ile Asp Ala Ser Val Val Glu Tyr Ser Lys Ile Ala Ile Gly Gln  
100 105 110  
Thr Asp Trp Asp Tyr Asn Arg Val Ala Ala Val Gln Glu Lys Tyr Met  
115 120 125  
Glu Asp Glu Glu Ile Gln Asp  
130 135

(2) INFORMATION FOR SEQ ID NO:577:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 518 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..518
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482457

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:577:

gattttgaga aaaccatctc actgattagc caagatgtcg tcggctcggt ctgcgatcac 60  
aaagctaaag ctggctcgat cctttgggga gagtcagatt ggtgcatcgc gttcgggtggt 120  
atcgactcga ggaccggcga ttcgggtactt cagtgcagat aaaggtcgtg tgctcagcga 180  
agaggaacgc gcgaaagaga gcatgtatat ccagaaaatg gagagggaaa gactggagaa 240  
gaagaagaaa ctcgagcaag ataagctaga tggtagagaaa ggaagtgcc acaagaaacc 300  
tgagacaagc aagccatgag tttatcactc acagtataca gaatccggtc ataaggcaag 360  
cagtagtgaa aaacaataat gcctttgacc tatgttctct cttggtatga gagatcttgt 420  
acttgtagcag agatcttttaa ccttctgtatg tgtgtgtttg tatgttctaa gaaatcaagt 480  
ttaaataaat cgaaaaaac aaccatatgc gttgattc

(2) INFORMATION FOR SEQ ID NO:578:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 51 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide  
(B) LOCATION: 1..51  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482458

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:578:

```
Asp Phe Glu Lys Thr Ile Ser Leu Ile Ser Gln Asp Val Val Gly Ser
1 5 10 15
Phe Cys Asp His Lys Ala Lys Ala Gly Ser Ile Leu Trp Gly Glu Ser
 20 25 30
Asp Trp Cys Ile Ala Phe Gly Gly Ile Asp Ser Arg Thr Gly Asp Ser
 35 40 45
Val Leu Gln
50
```

(2) INFORMATION FOR SEQ ID NO:579:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 94 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS:  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide  
(B) LOCATION: 1..94  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482459

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:579:

```
Met Ser Ser Ala Arg Ser Ala Ile Thr Lys Leu Lys Leu Ala Arg Ser
1 5 10 15
Phe Gly Glu Ser Gln Ile Gly Ala Ser Arg Ser Val Val Ser Thr Arg
 20 25 30
Gly Pro Ala Ile Arg Tyr Phe Ser Asp Asp Lys Gly Arg Val Leu Ser
 35 40 45
Glu Glu Glu Arg Ala Lys Glu Ser Met Tyr Ile Gln Lys Met Glu Arg
 50 55 60
Glu Arg Leu Glu Lys Lys Lys Lys Leu Glu Gln Asp Lys Leu Asp Gly
 65 70 75 80
Glu Lys Gly Ser Ala Asn Lys Lys Pro Glu Thr Ser Lys Pro
 85 90
```

(2) INFORMATION FOR SEQ ID NO:580:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 1116 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

(A) NAME/KEY: -  
(B) LOCATION: 1..1116  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482460

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:580:

```
atagaacatc ctaatcgaaa aacattagtt ttgctgctgt tagtattcaa ttcatcgcac 60
ccaaatcaaa atatatagga tactagataa agagtgcactg aaggagagaa aaacaaaaat 120
ggcgattggt tccgtctcta actcttttct cactttcaat tctccaatc agctccgatt 180
tagccgaaga agattctctg ccatggcttc ttcaactact ggagtgcgag tcgctgaagg 240
agaaggcaat ttgccaaaac tagtccttac ttctcctcag aacagcgagg ctgagatata 300
tctcttcgga ggctgcatta cttcttgga agttgcgagc ggtaaagatc ttctttttgt 360
cagaccagat gctgtcttca ataagattaa gccattagc ggagggattc cacattgttt 420
tccacagttt ggacctgggc taattcaaca gcatgggttt ggaaggaaca tggactggtc 480
tggtgtcgat tcccagaatg cagatgacaa tgctgctgtt actcttgagc ttaaggatgg 540
tccctatagt cgagccatgt gggactttgc tttccaggct ctatacaagg tcattgttgg 600
cgcggactcc ctttcactg agctaaagat tacaacacac gacgataaac cattttcttt 660
```

```
cagcactgcg ctgcatactt acttccgtgc ttcttctgcg ggggcctccg tgagaggtct 720
aaaggggtgt aaaaccctca ataaggatcc agaccctaag aaccaatag agggtaaaga 780
agacagggat gcagtcactt ttcctggatt tgtggatacc gtctatcttg atgctcccaa 840
tgaattgcag ttgtataatg gcttgggtga taaaataatc atcaaaaaca caaattggtc 900
ggatgcggtc ttgtggaacc cgcatactca gatggaggct tgttacagag actttgtgtg 960
cgtggaaaat gcaaagcttg gggatgtcaa gctagagccg ggacagtctt ggactgcaac 1020
acaacttctc agcatcagtt gaaaacattg tactttaaac ttataatgtc cagtggatcc 1080
atattcttaa gcaataaaaag ttttatttcc tctccc
```

(2) INFORMATION FOR SEQ ID NO:581:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 307 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..307

(D) OTHER INFORMATION: / Ceres Seq. ID 1482461

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:581:

```
Met Ala Ile Val Ser Val Ser Asn Ser Phe Leu Thr Phe Asn Ser Pro
1 5 10 15
Asn Gln Leu Arg Phe Ser Arg Arg Arg Phe Ser Ala Met Ala Ser Ser
 20 25 30
Thr Thr Gly Val Arg Val Ala Glu Gly Glu Gly Asn Leu Pro Lys Leu
 35 40 45
Val Leu Thr Ser Pro Gln Asn Ser Glu Ala Glu Ile Tyr Leu Phe Gly
 50 55 60
Gly Cys Ile Thr Ser Trp Lys Val Ala Ser Gly Lys Asp Leu Leu Phe
 65 70 75 80
Val Arg Pro Asp Ala Val Phe Asn Lys Ile Lys Pro Ile Ser Gly Gly
 85 90 95
Ile Pro His Cys Phe Pro Gln Phe Gly Pro Gly Leu Ile Gln Gln His
 100 105 110
Gly Phe Gly Arg Asn Met Asp Trp Ser Val Val Asp Ser Gln Asn Ala
 115 120 125
Asp Asp Asn Ala Ala Val Thr Leu Glu Leu Lys Asp Gly Pro Tyr Ser
 130 135 140
Arg Ala Met Trp Asp Phe Ala Phe Gln Ala Leu Tyr Lys Val Ile Val
 145 150 155 160
Gly Ala Asp Ser Leu Ser Thr Glu Leu Lys Ile Thr Asn Thr Asp Asp
 165 170 175
Lys Pro Phe Ser Phe Ser Thr Ala Leu His Thr Tyr Phe Arg Ala Ser
 180 185 190
Ser Ala Gly Ala Ser Val Arg Gly Leu Lys Gly Cys Lys Thr Leu Asn
 195 200 205
Lys Asp Pro Asp Pro Lys Asn Pro Ile Glu Gly Lys Glu Asp Arg Asp
 210 215 220
Ala Val Thr Phe Pro Gly Phe Val Asp Thr Val Tyr Leu Asp Ala Pro
 225 230 235 240
Asn Glu Leu Gln Phe Asp Asn Gly Leu Gly Asp Lys Ile Ile Ile Lys
 245 250 255
Asn Thr Asn Trp Ser Asp Ala Val Leu Trp Asn Pro His Thr Gln Met
 260 265 270
Glu Ala Cys Tyr Arg Asp Phe Val Cys Val Glu Asn Ala Lys Leu Gly
 275 280 285
Asp Val Lys Leu Glu Pro Gly Gln Ser Trp Thr Ala Thr Gln Leu Leu
 290 295 300
Ser Ile Ser
```

305

(2) INFORMATION FOR SEQ ID NO:582:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 279 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..279
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482462

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:582:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |  |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|--|
| Met | Ala | Ser | Ser | Thr | Thr | Gly | Val | Arg | Val | Ala | Glu | Gly | Glu | Gly | Asn |  |
| 1   |     |     |     | 5   |     |     |     |     | 10  |     |     |     | 15  |     |     |  |
| Leu | Pro | Lys | Leu | Val | Leu | Thr | Ser | Pro | Gln | Asn | Ser | Glu | Ala | Glu | Ile |  |
|     |     |     | 20  |     |     |     |     | 25  |     |     |     | 30  |     |     |     |  |
| Tyr | Leu | Phe | Gly | Gly | Cys | Ile | Thr | Ser | Trp | Lys | Val | Ala | Ser | Gly | Lys |  |
|     |     | 35  |     |     |     | 40  |     |     |     |     | 45  |     |     |     |     |  |
| Asp | Leu | Leu | Phe | Val | Arg | Pro | Asp | Ala | Val | Phe | Asn | Lys | Ile | Lys | Pro |  |
|     | 50  |     |     |     | 55  |     |     |     |     | 60  |     |     |     |     |     |  |
| Ile | Ser | Gly | Gly | Ile | Pro | His | Cys | Phe | Pro | Gln | Phe | Gly | Pro | Gly | Leu |  |
| 65  |     |     |     |     | 70  |     |     |     |     | 75  |     |     |     |     | 80  |  |
| Ile | Gln | Gln | His | Gly | Phe | Gly | Arg | Asn | Met | Asp | Trp | Ser | Val | Val | Asp |  |
|     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |     |  |
| Ser | Gln | Asn | Ala | Asp | Asp | Asn | Ala | Ala | Val | Thr | Leu | Glu | Leu | Lys | Asp |  |
|     |     |     | 100 |     |     |     |     | 105 |     |     |     |     | 110 |     |     |  |
| Gly | Pro | Tyr | Ser | Arg | Ala | Met | Trp | Asp | Phe | Ala | Phe | Gln | Ala | Leu | Tyr |  |
|     |     | 115 |     |     |     | 120 |     |     |     |     |     | 125 |     |     |     |  |
| Lys | Val | Ile | Val | Gly | Ala | Asp | Ser | Leu | Ser | Thr | Glu | Leu | Lys | Ile | Thr |  |
|     | 130 |     |     |     | 135 |     |     |     |     |     | 140 |     |     |     |     |  |
| Asn | Thr | Asp | Asp | Lys | Pro | Phe | Ser | Phe | Ser | Thr | Ala | Leu | His | Thr | Tyr |  |
| 145 |     |     |     |     | 150 |     |     |     |     | 155 |     |     |     |     | 160 |  |
| Phe | Arg | Ala | Ser | Ser | Ala | Gly | Ala | Ser | Val | Arg | Gly | Leu | Lys | Gly | Cys |  |
|     |     |     | 165 |     |     |     |     | 170 |     |     |     |     |     | 175 |     |  |
| Lys | Thr | Leu | Asn | Lys | Asp | Pro | Asp | Pro | Lys | Asn | Pro | Ile | Glu | Gly | Lys |  |
|     |     |     | 180 |     |     |     |     | 185 |     |     |     |     | 190 |     |     |  |
| Glu | Asp | Arg | Asp | Ala | Val | Thr | Phe | Pro | Gly | Phe | Val | Asp | Thr | Val | Tyr |  |
|     | 195 |     |     |     |     | 200 |     |     |     |     |     | 205 |     |     |     |  |
| Leu | Asp | Ala | Pro | Asn | Glu | Leu | Gln | Phe | Asp | Asn | Gly | Leu | Gly | Asp | Lys |  |
|     | 210 |     |     |     |     | 215 |     |     |     |     | 220 |     |     |     |     |  |
| Ile | Ile | Ile | Lys | Asn | Thr | Asn | Trp | Ser | Asp | Ala | Val | Leu | Trp | Asn | Pro |  |
| 225 |     |     |     |     | 230 |     |     |     |     | 235 |     |     |     |     | 240 |  |
| His | Thr | Gln | Met | Glu | Ala | Cys | Tyr | Arg | Asp | Phe | Val | Cys | Val | Glu | Asn |  |
|     |     |     | 245 |     |     |     |     | 250 |     |     |     |     |     | 255 |     |  |
| Ala | Lys | Leu | Gly | Asp | Val | Lys | Leu | Glu | Pro | Gly | Gln | Ser | Trp | Thr | Ala |  |
|     |     | 260 |     |     |     |     |     | 265 |     |     |     |     |     | 270 |     |  |
| Thr | Gln | Leu | Leu | Ser | Ile | Ser |     |     |     |     |     |     |     |     |     |  |
|     |     |     | 275 |     |     |     |     |     |     |     |     |     |     |     |     |  |

(2) INFORMATION FOR SEQ ID NO:583:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 190 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..190



(D) OTHER INFORMATION: / Ceres Seq. ID 1482463

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:583:

```
Met Asp Trp Ser Val Val Asp Ser Gln Asn Ala Asp Asp Asn Ala Ala
1 5 10 15
Val Thr Leu Glu Leu Lys Asp Gly Pro Tyr Ser Arg Ala Met Trp Asp
 20 25 30
Phe Ala Phe Gln Ala Leu Tyr Lys Val Ile Val Gly Ala Asp Ser Leu
 35 40 45
Ser Thr Glu Leu Lys Ile Thr Asn Thr Asp Asp Lys Pro Phe Ser Phe
 50 55 60
Ser Thr Ala Leu His Thr Tyr Phe Arg Ala Ser Ser Ala Gly Ala Ser
 65 70 75 80
Val Arg Gly Leu Lys Gly Cys Lys Thr Leu Asn Lys Asp Pro Asp Pro
 85 90 95
Lys Asn Pro Ile Glu Gly Lys Glu Asp Arg Asp Ala Val Thr Phe Pro
 100 105 110
Gly Phe Val Asp Thr Val Tyr Leu Asp Ala Pro Asn Glu Leu Gln Phe
 115 120 125
Asp Asn Gly Leu Gly Asp Lys Ile Ile Ile Lys Asn Thr Asn Trp Ser
 130 135 140
Asp Ala Val Leu Trp Asn Pro His Thr Gln Met Glu Ala Cys Tyr Arg
 145 150 155 160
Asp Phe Val Cys Val Glu Asn Ala Lys Leu Gly Asp Val Lys Leu Glu
 165 170 175
Pro Gly Gln Ser Trp Thr Ala Thr Gln Leu Leu Ser Ile Ser
 180 185 190
```

(2) INFORMATION FOR SEQ ID NO:584:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1430 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1430

(D) OTHER INFORMATION: / Ceres Seq. ID 1482481

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:584:

```
ccwagaamca tcctaatcaa aaaacaattc ccgaaattct ctcaaatcac agatcctttt 60
agggtttttc cactgtttct aggttttttt tattgctcaa atctgatcaa tggatagttg 120
tctctctaata caaacggcgc ttcagttttt cccgtcgcgt tccaggagac agagcggcga 180
tggaggcgggt ggttttggtta ttccggcgaa gaggaaagac cagtatagtt cgtggttgt 240
ggttgcggcg gcgggacaga gtcgggtgta gcctggaagc agtctaaacg cgcgcttga 300
gccacgatcg gcgcagggga ggtttctgag aagcgtgttg ctaaacaacac ggcagctatt 360
tcattacgcc gccgtgatg agctaaagca actggctgat gatagggaag ctgcttttagc 420
tcgtatgtct ctacagctctg gttccgatga ggcttctctc cacagaagga tagctgaact 480
caaggaacgc tactgtaaaa ctgcagtcca agacataatg tacatgttaa tcttttacaa 540
atactccgag ataagagtcc ctcttggttc aaagctatcc agatgcatct ataattggaag 600
actcgagatc tggcctttcaa aagactggga gttagagtca atttacagct gcgataccct 660
tgagatcatc aaagaacacg ttagcgcagt catcggatta cgggtcaact catgtgtgac 720
tgacaattgg gcaacaacgc agatacagaa actgcatctc aggaaagtat atgctgcctc 780
gatcttgtag ggttacttct tgaaatcagc ttccctaagg caccagcttg agtgttcctt 840
atcagatatt catggaagcg gatattctgaa aagtcccatc tttggatgct cattcacaac 900
gggcactgca cagatctcca acaagcagca gctgagacat tacatctcag actttgatcc 960
cgagacattg cagagatgcg caaaaccaag gacagaggag gcaaggaatc tgatagagaa 1020
gcaaagtttg gctotttttg gcacggaaga gagtgatgag accatagtga catcgttttc 1080
gagtctgaag cggttggttc tcgaggctgt ggcgtttggg acattcctgt gggacacgga 1140
attgtatgta gatggtgcat ataagctgaa ggagaatggg aatgcagaag aacaagaagg 1200
aaagaaaagc atatgatgaa caagtctggt tagaagaaaa gcttcatgat cttctggtag 1260
```

tgtatatata gagaaatgta tctgccgaat ctctcaggca gttgttcagt tcaatgtata 1320  
gatcttgctt agaaatattt tgatttctga ataagaatgt ggtgtgggta taaggaataa 1380  
gagatactgt agttgggttc aattttatgt tatgtgttaa gtttccttgt

(2) INFORMATION FOR SEQ ID NO:585:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 368 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..368

(D) OTHER INFORMATION: / Ceres Seq. ID 1482482

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:585:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Asp | Ser | Cys | Leu | Ser | Asn | Gln | Thr | Ala | Leu | Gln | Phe | Leu | Pro | Ser |
| 1   |     |     | 5   |     |     |     | 10  |     |     |     | 15  |     |     |     |     |
| Arg | Ser | Arg | Arg | Gln | Ser | Gly | Asp | Gly | Gly | Gly | Gly | Phe | Val | Ile | Pro |
|     |     |     | 20  |     |     |     | 25  |     |     |     | 30  |     |     |     |     |
| Ala | Lys | Arg | Lys | Ile | Gln | Tyr | Ser | Ser | Met | Val | Val | Val | Ala | Ala | Ala |
|     |     |     | 35  |     |     |     | 40  |     |     |     | 45  |     |     |     |     |
| Gly | Gln | Ser | Arg | Cys | Glu | Pro | Gly | Ser | Ser | Leu | Asn | Ala | Pro | Leu | Glu |
|     |     |     | 50  |     |     |     | 55  |     |     |     | 60  |     |     |     |     |
| Pro | Arg | Ser | Ala | Gln | Gly | Arg | Phe | Leu | Arg | Ser | Val | Leu | Leu | Asn | Lys |
| 65  |     |     |     |     |     | 70  |     |     |     | 75  |     |     |     | 80  |     |
| Arg | Gln | Leu | Phe | His | Tyr | Ala | Ala | Ala | Asp | Glu | Leu | Lys | Gln | Leu | Ala |
|     |     |     |     |     |     | 85  |     |     |     | 90  |     |     |     | 95  |     |
| Asp | Asp | Arg | Glu | Ala | Ala | Leu | Ala | Arg | Met | Ser | Leu | Ser | Ser | Gly | Ser |
|     |     |     | 100 |     |     |     |     |     | 105 |     |     |     | 110 |     |     |
| Asp | Glu | Ala | Ser | Leu | His | Arg | Arg | Ile | Ala | Glu | Leu | Lys | Glu | Arg | Tyr |
|     |     |     | 115 |     |     |     |     |     | 120 |     |     |     | 125 |     |     |
| Cys | Lys | Thr | Ala | Val | Gln | Asp | Ile | Met | Tyr | Met | Leu | Ile | Phe | Tyr | Lys |
|     |     |     | 130 |     |     |     | 135 |     |     |     | 140 |     |     |     |     |
| Tyr | Ser | Glu | Ile | Arg | Val | Pro | Leu | Val | Pro | Lys | Leu | Ser | Arg | Cys | Ile |
| 145 |     |     |     |     |     | 150 |     |     |     | 155 |     |     |     | 160 |     |
| Tyr | Asn | Gly | Arg | Leu | Glu | Ile | Trp | Pro | Ser | Lys | Asp | Trp | Glu | Leu | Glu |
|     |     |     |     |     |     | 165 |     |     |     | 170 |     |     |     | 175 |     |
| Ser | Ile | Tyr | Ser | Cys | Asp | Thr | Leu | Glu | Ile | Ile | Lys | Glu | His | Val | Ser |
|     |     |     |     |     |     | 180 |     |     |     | 185 |     |     |     | 190 |     |
| Ala | Val | Ile | Gly | Leu | Arg | Val | Asn | Ser | Cys | Val | Thr | Asp | Asn | Trp | Ala |
|     |     |     | 195 |     |     |     | 200 |     |     |     |     |     | 205 |     |     |
| Thr | Thr | Gln | Ile | Gln | Lys | Leu | His | Leu | Arg | Lys | Val | Tyr | Ala | Ala | Ser |
|     |     |     |     |     |     | 210 |     |     |     | 215 |     |     | 220 |     |     |
| Ile | Leu | Tyr | Gly | Tyr | Phe | Leu | Lys | Ser | Ala | Ser | Leu | Arg | His | Gln | Leu |
| 225 |     |     |     |     |     | 230 |     |     |     | 235 |     |     |     | 240 |     |
| Glu | Cys | Ser | Leu | Ser | Asp | Ile | His | Gly | Ser | Gly | Tyr | Leu | Lys | Ser | Pro |
|     |     |     |     |     |     | 245 |     |     |     | 250 |     |     |     | 255 |     |
| Ile | Phe | Gly | Cys | Ser | Phe | Thr | Thr | Gly | Thr | Ala | Gln | Ile | Ser | Asn | Lys |
|     |     |     | 260 |     |     |     |     |     | 265 |     |     |     |     | 270 |     |
| Gln | Gln | Leu | Arg | His | Tyr | Ile | Ser | Asp | Phe | Asp | Pro | Glu | Thr | Leu | Gln |
|     |     |     | 275 |     |     |     |     |     | 280 |     |     |     | 285 |     |     |
| Arg | Cys | Ala | Lys | Pro | Arg | Thr | Glu | Glu | Ala | Arg | Asn | Leu | Ile | Glu | Lys |
|     |     |     | 290 |     |     |     | 295 |     |     |     | 300 |     |     |     |     |
| Gln | Ser | Leu | Ala | Leu | Phe | Gly | Thr | Glu | Glu | Ser | Asp | Glu | Thr | Ile | Val |
| 305 |     |     |     |     |     | 310 |     |     |     | 315 |     |     |     | 320 |     |
| Thr | Ser | Phe | Ser | Ser | Leu | Lys | Arg | Leu | Val | Leu | Glu | Ala | Val | Ala | Phe |
|     |     |     |     |     |     | 325 |     |     |     | 330 |     |     |     | 335 |     |
| Gly | Thr | Phe | Leu | Trp | Asp | Thr | Glu | Leu | Tyr | Val | Asp | Gly | Ala | Tyr | Lys |
|     |     |     |     |     |     | 340 |     |     |     | 345 |     |     |     | 350 |     |

Leu Lys Glu Asn Gly Asn Ala Glu Glu Gln Glu Gly Lys Lys Ser Ile  
355 360 365

(2) INFORMATION FOR SEQ ID NO:586:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 327 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..327

(D) OTHER INFORMATION: / Ceres Seq. ID 1482483

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:586:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |  |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|--|
| Met | Val | Val | Val | Ala | Ala | Ala | Gly | Gln | Ser | Arg | Cys | Glu | Pro | Gly | Ser |  |
| 1   |     |     |     | 5   |     |     |     | 10  |     |     |     |     |     | 15  |     |  |
| Ser | Leu | Asn | Ala | Pro | Leu | Glu | Pro | Arg | Ser | Ala | Gln | Gly | Arg | Phe | Leu |  |
|     |     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |  |
| Arg | Ser | Val | Leu | Leu | Asn | Lys | Arg | Gln | Leu | Phe | His | Tyr | Ala | Ala | Ala |  |
|     |     |     | 35  |     |     |     | 40  |     |     |     |     | 45  |     |     |     |  |
| Asp | Glu | Leu | Lys | Gln | Leu | Ala | Asp | Asp | Arg | Glu | Ala | Ala | Leu | Ala | Arg |  |
|     | 50  |     |     |     |     | 55  |     |     |     |     | 60  |     |     |     |     |  |
| Met | Ser | Leu | Ser | Ser | Gly | Ser | Asp | Glu | Ala | Ser | Leu | His | Arg | Arg | Ile |  |
|     |     |     |     |     | 70  |     |     |     |     |     | 75  |     |     |     | 80  |  |
| Ala | Glu | Leu | Lys | Glu | Arg | Tyr | Cys | Lys | Thr | Ala | Val | Gln | Asp | Ile | Met |  |
|     |     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |  |
| Tyr | Met | Leu | Ile | Phe | Tyr | Lys | Tyr | Ser | Glu | Ile | Arg | Val | Pro | Leu | Val |  |
|     |     |     | 100 |     |     |     |     | 105 |     |     |     |     | 110 |     |     |  |
| Pro | Lys | Leu | Ser | Arg | Cys | Ile | Tyr | Asn | Gly | Arg | Leu | Glu | Ile | Trp | Pro |  |
|     |     |     | 115 |     |     |     |     | 120 |     |     |     |     | 125 |     |     |  |
| Ser | Lys | Asp | Trp | Glu | Leu | Glu | Ser | Ile | Tyr | Ser | Cys | Asp | Thr | Leu | Glu |  |
|     |     |     |     |     |     | 135 |     |     |     |     | 140 |     |     |     |     |  |
| Ile | Ile | Lys | Glu | His | Val | Ser | Ala | Val | Ile | Gly | Leu | Arg | Val | Asn | Ser |  |
|     |     |     |     | 150 |     |     |     |     |     | 155 |     |     |     | 160 |     |  |
| Cys | Val | Thr | Asp | Asn | Trp | Ala | Thr | Thr | Gln | Ile | Gln | Lys | Leu | His | Leu |  |
|     |     |     |     | 165 |     |     |     |     | 170 |     |     |     |     | 175 |     |  |
| Arg | Lys | Val | Tyr | Ala | Ala | Ser | Ile | Leu | Tyr | Gly | Tyr | Phe | Leu | Lys | Ser |  |
|     |     |     | 180 |     |     |     |     | 185 |     |     |     |     | 190 |     |     |  |
| Ala | Ser | Leu | Arg | His | Gln | Leu | Glu | Cys | Ser | Leu | Ser | Asp | Ile | His | Gly |  |
|     |     |     | 195 |     |     |     | 200 |     |     |     |     | 205 |     |     |     |  |
| Ser | Gly | Tyr | Leu | Lys | Ser | Pro | Ile | Phe | Gly | Cys | Ser | Phe | Thr | Thr | Gly |  |
|     |     |     | 210 |     |     |     | 215 |     |     |     |     | 220 |     |     |     |  |
| Thr | Ala | Gln | Ile | Ser | Asn | Lys | Gln | Gln | Leu | Arg | His | Tyr | Ile | Ser | Asp |  |
|     |     |     |     | 230 |     |     |     |     |     | 235 |     |     |     | 240 |     |  |
| Phe | Asp | Pro | Glu | Thr | Leu | Gln | Arg | Cys | Ala | Lys | Pro | Arg | Thr | Glu | Glu |  |
|     |     |     |     | 245 |     |     |     |     | 250 |     |     |     |     | 255 |     |  |
| Ala | Arg | Asn | Leu | Ile | Glu | Lys | Gln | Ser | Leu | Ala | Leu | Phe | Gly | Thr | Glu |  |
|     |     |     | 260 |     |     |     | 265 |     |     |     |     |     | 270 |     |     |  |
| Glu | Ser | Asp | Glu | Thr | Ile | Val | Thr | Ser | Phe | Ser | Ser | Leu | Lys | Arg | Leu |  |
|     |     |     | 275 |     |     |     | 280 |     |     |     |     | 285 |     |     |     |  |
| Val | Leu | Glu | Ala | Val | Ala | Phe | Gly | Thr | Phe | Leu | Trp | Asp | Thr | Glu | Leu |  |
|     |     |     | 290 |     |     | 295 |     |     |     |     | 300 |     |     |     |     |  |
| Tyr | Val | Asp | Gly | Ala | Tyr | Lys | Leu | Lys | Glu | Asn | Gly | Asn | Ala | Glu | Glu |  |
|     |     |     | 305 |     |     | 310 |     |     |     | 315 |     |     |     | 320 |     |  |
| Gln | Glu | Gly | Lys | Lys | Ser | Ile |     |     |     |     |     |     |     |     |     |  |
|     |     |     |     | 325 |     |     |     |     |     |     |     |     |     |     |     |  |

(2) INFORMATION FOR SEQ ID NO:587:

(i) SEQUENCE CHARACTERISTICS:  
    (A) LENGTH: 263 amino acids  
    (B) TYPE: amino acid  
    (C) STRANDEDNESS:  
    (D) TOPOLOGY: linear  
(ii) MOLECULE TYPE: peptide  
(ix) FEATURE:  
    (A) NAME/KEY: peptide  
    (B) LOCATION: 1..263  
    (D) OTHER INFORMATION: / Ceres Seq. ID 1482484  
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:587:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Ser | Leu | Ser | Ser | Gly | Ser | Asp | Glu | Ala | Ser | Leu | His | Arg | Arg | Ile |
| 1   |     |     |     | 5   |     |     |     |     | 10  |     |     |     |     | 15  |     |
| Ala | Glu | Leu | Lys | Glu | Arg | Tyr | Cys | Lys | Thr | Ala | Val | Gln | Asp | Ile | Met |
|     |     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |
| Tyr | Met | Leu | Ile | Phe | Tyr | Lys | Tyr | Ser | Glu | Ile | Arg | Val | Pro | Leu | Val |
|     |     | 35  |     |     |     | 40  |     |     |     |     | 45  |     |     |     |     |
| Pro | Lys | Leu | Ser | Arg | Cys | Ile | Tyr | Asn | Gly | Arg | Leu | Glu | Ile | Trp | Pro |
|     | 50  |     |     |     | 55  |     |     |     | 60  |     |     |     |     |     |     |
| Ser | Lys | Asp | Trp | Glu | Leu | Glu | Ser | Ile | Tyr | Ser | Cys | Asp | Thr | Leu | Glu |
| 65  |     |     |     | 70  |     |     |     | 75  |     |     |     |     |     | 80  |     |
| Ile | Ile | Lys | Glu | His | Val | Ser | Ala | Val | Ile | Gly | Leu | Arg | Val | Asn | Ser |
|     |     |     | 85  |     |     |     |     | 90  |     |     |     |     |     | 95  |     |
| Cys | Val | Thr | Asp | Asn | Trp | Ala | Thr | Thr | Gln | Ile | Gln | Lys | Leu | His | Leu |
|     |     |     | 100 |     |     |     |     | 105 |     |     |     |     | 110 |     |     |
| Arg | Lys | Val | Tyr | Ala | Ala | Ser | Ile | Leu | Tyr | Gly | Tyr | Phe | Leu | Lys | Ser |
|     |     |     | 115 |     |     |     | 120 |     |     |     |     | 125 |     |     |     |
| Ala | Ser | Leu | Arg | His | Gln | Leu | Glu | Cys | Ser | Leu | Ser | Asp | Ile | His | Gly |
|     |     |     | 130 |     |     |     | 135 |     |     |     |     | 140 |     |     |     |
| Ser | Gly | Tyr | Leu | Lys | Ser | Pro | Ile | Phe | Gly | Cys | Ser | Phe | Thr | Thr | Gly |
| 145 |     |     |     | 150 |     |     |     |     | 155 |     |     |     |     | 160 |     |
| Thr | Ala | Gln | Ile | Ser | Asn | Lys | Gln | Gln | Leu | Arg | His | Tyr | Ile | Ser | Asp |
|     |     |     | 165 |     |     |     |     | 170 |     |     |     |     |     | 175 |     |
| Phe | Asp | Pro | Glu | Thr | Leu | Gln | Arg | Cys | Ala | Lys | Pro | Arg | Thr | Glu | Glu |
|     |     |     | 180 |     |     |     |     | 185 |     |     |     |     | 190 |     |     |
| Ala | Arg | Asn | Leu | Ile | Glu | Lys | Gln | Ser | Leu | Ala | Leu | Phe | Gly | Thr | Glu |
|     |     |     | 195 |     |     |     | 200 |     |     |     |     | 205 |     |     |     |
| Glu | Ser | Asp | Glu | Thr | Ile | Val | Thr | Ser | Phe | Ser | Ser | Leu | Lys | Arg | Leu |
|     |     |     | 210 |     |     | 215 |     |     |     |     | 220 |     |     |     |     |
| Val | Leu | Glu | Ala | Val | Ala | Phe | Gly | Thr | Phe | Leu | Trp | Asp | Thr | Glu | Leu |
| 225 |     |     |     | 230 |     |     |     |     | 235 |     |     |     |     | 240 |     |
| Tyr | Val | Asp | Gly | Ala | Tyr | Lys | Leu | Lys | Glu | Asn | Gly | Asn | Ala | Glu | Glu |
|     |     |     | 245 |     |     |     |     | 250 |     |     |     |     |     | 255 |     |
| Gln | Glu | Gly | Lys | Lys | Ser | Ile |     |     |     |     |     |     |     |     |     |
|     |     |     |     | 260 |     |     |     |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:588:

(i) SEQUENCE CHARACTERISTICS:  
    (A) LENGTH: 662 base pairs  
    (B) TYPE: nucleic acid  
    (C) STRANDEDNESS: single  
    (D) TOPOLOGY: linear  
(ii) MOLECULE TYPE: DNA (genomic)  
(ix) FEATURE:  
    (A) NAME/KEY: -  
    (B) LOCATION: 1..662  
    (D) OTHER INFORMATION: / Ceres Seq. ID 1482490  
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:588:

|            |            |            |            |            |            |     |
|------------|------------|------------|------------|------------|------------|-----|
| atcgaaaaag | atcgaaaaaa | aatcgagaag | cgaatagcgg | aagaacagaa | aaagggaaat | 60  |
| tctgagaatc | aaatcgaaaa | ggtagaagaa | tcgagtcgga | aaaatggaaa | cgacgaaaag | 120 |

```
taacagcagc gagtccgatg tcaacgccaa atgggacgct tgtctcgatc tcactgctcg 180
tcgctttgtc tactcttccc tcggcgggcg tttcgccggt cttctcttct tcaggagtcc 240
ggttacgaga tgggcgtcga ttgcttttgg tgctggaatt ggtattggtt ctgcatacac 300
agattgttct cgtgtttttg atgcgtcttc ttcaacttca gctactttat tagcagctcc 360
caagagtaca gagacttctg tatctcaggc agcagaagag tgaagacaac gaggaagctt 420
ggaggtaaaa aaccaaacat tgataggggt acattacgaa atggtaattg atcttgccagg 480
acaaggcttt tgagataacg ccattgttaa aaaaaaactt ttgcttctca gtgtggtttt 540
gtacactgat gtcaaaattg ttaatgaccc actcattttt ttttgttttg aaaaatctta 600
tgttctttta cttgagaaat aattcctccg ttgatttggt tgcctctact gttccttcat 660
tc
```

(2) INFORMATION FOR SEQ ID NO:589:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 99 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..99
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482491

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:589:

```
Met Glu Thr Thr Lys Ser Asn Ser Ser Glu Ser Asp Val Asn Ala Lys
1 5 10 15
Trp Asp Ala Cys Leu Asp Leu Thr Ala Arg Arg Phe Val Tyr Ser Ser
 20 25 30
Leu Gly Gly Ala Phe Ala Gly Leu Leu Phe Phe Arg Ser Pro Val Thr
 35 40 45
Arg Trp Ala Ser Ile Ala Phe Gly Ala Gly Ile Gly Ile Gly Ser Ala
 50 55 60
Tyr Thr Asp Cys Ser Arg Val Phe Asp Ala Ser Ser Ser Thr Ser Ala
65 70 75 80
Thr Leu Leu Ala Ala Pro Lys Ser Thr Glu Thr Ser Val Ser Gln Ala
 85 90 95
Ala Glu Glu
```

(2) INFORMATION FOR SEQ ID NO:590:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 71 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..71
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482492

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:590:

```
Met Ser Thr Pro Asn Gly Thr Leu Val Ser Ile Ser Leu Leu Val Ala
1 5 10 15
Leu Ser Thr Leu Pro Ser Ala Ala Leu Ser Pro Val Phe Ser Ser Ser
 20 25 30
Gly Val Arg Leu Arg Asp Gly Arg Arg Leu Leu Leu Val Leu Glu Leu
 35 40 45
Val Leu Val Leu His Thr Gln Ile Val Leu Val Phe Leu Met Arg Leu
 50 55 60
Leu Gln Leu Gln Leu Leu Tyr
65 70
```

(2) INFORMATION FOR SEQ ID NO:591:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 56 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..56
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482493

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:591:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Gly | Arg | Leu | Ser | Arg | Ser | His | Cys | Ser | Ser | Leu | Cys | Leu | Leu | Phe |
| 1   |     |     |     | 5   |     |     |     | 10  |     |     |     |     | 15  |     |     |
| Pro | Arg | Arg | Arg | Phe | Arg | Arg | Ser | Ser | Leu | Leu | Gln | Glu | Ser | Gly | Tyr |
|     |     |     | 20  |     |     |     | 25  |     |     |     |     | 30  |     |     |     |
| Glu | Met | Gly | Val | Asp | Cys | Phe | Trp | Cys | Trp | Asn | Trp | Tyr | Trp | Phe | Cys |
|     |     | 35  |     |     |     |     | 40  |     |     |     |     | 45  |     |     |     |
| Ile | His | Arg | Leu | Phe | Ser | Cys | Phe |     |     |     |     |     |     |     |     |
|     | 50  |     |     |     |     | 55  |     |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:592:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 853 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..853
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482504

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:592:

|            |            |             |            |            |            |     |
|------------|------------|-------------|------------|------------|------------|-----|
| ccattaccwa | kaacatccta | atcgaaaagt  | aatcggagtt | caggcttcag | cattctctct | 60  |
| tcttctctct | cgcagccgta | gtttttgata  | ttctcttcaa | ttctctctcc | tgatggccac | 120 |
| gtccgcccgc | ctctccggtg | ccagatcgat  | gcttcgagct | gcttcctcac | gcagcgccgc | 180 |
| tgcttctact | ggccgcttcg | cctctcaagc  | gaaatccgct | ccaccattgt | ttagagccac | 240 |
| tgccagaaga | agcccactgc | tttctcctct  | ccgaaatcct | gtggaactga | gcttctgtgt | 300 |
| ggagtcattg | ttaccatata | actcggctac  | agcttcagcg | ctaatactt  | caaagctttc | 360 |
| tatctctggc | caaacctatg | gctggctctc  | tgacggctga | cacaagtgt  | gatgaagaca | 420 |
| acgaagccaa | gatctgggta | taaacgatta  | gaacgggttt | caggcaataa | gataggcttt | 480 |
| agatacacat | caagcaatgg | ttgatgctgc  | atttgtttt  | aaaagaactg | gttcttacat | 540 |
| atcttcttaa | aaaaaataca | tgtacccgga  | aaagtgcctt | cttttcttgd | tggttatagc | 600 |
| atttgagtta | ttactgattg | gtcttatact  | cccagcttgc | aatgatgatg | tgtgatgagt | 660 |
| tagccagagg | aacaatgaag | ctacagtitta | tgtacaaaac | tctacctttt | aaagcctttc | 720 |
| ttcttaaaaa | acttaggaac | gaaaaccctc  | ttaattttgt | ttctgagttt | cttggagagc | 780 |
| ttttgtttgt | tttcagccta | ttaagtaaga  | catggtgtat | tggttggacg | agtaactgat | 840 |
| gtttgttata | att        |             |            |            |            |     |

(2) INFORMATION FOR SEQ ID NO:593:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 95 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..95
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482505

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:593:

Met Ala Thr Ser Ala Val Leu Ser Gly Ala Arg Ser Met Leu Arg Ala

```

1 5 10 15
Ala Ser Ser Arg Ser Ala Ala Ala Ser Thr Gly Arg Phe Ala Ser Gln
 20 25 30
Ala Lys Ser Ala Pro Pro Leu Phe Arg Ala Thr Ala Arg Arg Ser Pro
 35 40 45
Leu Leu Ser Pro Leu Arg Asn Pro Val Glu Leu Ser Phe Cys Val Glu
 50 55 60
Ser Leu Leu Pro Tyr His Ser Ala Thr Ala Ser Ala Leu Met Thr Ser
65 70 75 80
Lys Leu Ser Ile Ser Gly Gln Thr Tyr Gly Trp Leu Ser Asp Gly
 85 90 95
```

(2) INFORMATION FOR SEQ ID NO:594:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 83 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..83
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482506

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:594:

```

Met Leu Arg Ala Ala Ser Ser Arg Ser Ala Ala Ala Ser Thr Gly Arg
1 5 10 15
Phe Ala Ser Gln Ala Lys Ser Ala Pro Pro Leu Phe Arg Ala Thr Ala
 20 25 30
Arg Arg Ser Pro Leu Leu Ser Pro Leu Arg Asn Pro Val Glu Leu Ser
 35 40 45
Phe Cys Val Glu Ser Leu Leu Pro Tyr His Ser Ala Thr Ala Ser Ala
50 55 60
Leu Met Thr Ser Lys Leu Ser Ile Ser Gly Gln Thr Tyr Gly Trp Leu
65 70 75 80
Ser Asp Gly
```

(2) INFORMATION FOR SEQ ID NO:595:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 65 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..65
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482507

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:595:

```

Met Met Met Cys Asp Glu Leu Ala Arg Gly Thr Met Lys Leu Gln Phe
1 5 10 15
Met Tyr Lys Thr Leu Pro Phe Lys Ala Phe Leu Leu Lys Lys Leu Arg
 20 25 30
Asn Glu Asn Pro Leu Asn Phe Val Ser Glu Phe Leu Gly Glu Leu Leu
 35 40 45
Phe Val Phe Ser Leu Leu Ser Lys Thr Cys Cys Ile Gly Trp Thr Ser
50 55 60
```

Asn

65

(2) INFORMATION FOR SEQ ID NO:596:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 1139 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear  
(ii) MOLECULE TYPE: DNA (genomic)  
(ix) FEATURE:  
(A) NAME/KEY: -  
(B) LOCATION: 1..1139  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482508  
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:596:

|             |            |            |             |             |             |      |
|-------------|------------|------------|-------------|-------------|-------------|------|
| caatcatctg  | atctttccct | ctctcagcaa | tgcattgatct | tcgattttct  | atcagtgttc  | 60   |
| aaagctgaaa  | aaaatcgaac | tgggtctggt | gatttcttca  | ggctctaaaat | cagattagat  | 120  |
| tagagaagaa  | gaagaagaat | gctggaagct | gtagatagct  | caggagtggg  | gaatggagga  | 180  |
| ttcccgacga  | ttcagagctt | ttacggcgat | tgcagtagtg  | aagaagagtt  | atcgggtattg | 240  |
| ccacgtcata  | caaaagtggg | ggtcaccgga | aacaaccgga  | cgaaatcggg  | gcttggtggg  | 300  |
| cttcaagggtg | ttgtcaaaaa | agctgtcggg | ctcgggtggg  | ggcattgggt  | ggttttgaca  | 360  |
| aatggaatag  | aagtaaaagt | gcagaggaat | gcgcttagtg  | tccttgaacc  | tcctactgga  | 420  |
| aacgaagaag  | acgatgatct | tgatttctga | aacacacaga  | ggaatggctc  | tgatatgatt  | 480  |
| gtttcttttc  | cagcatctga | ggacacactg | aagcctcata  | agtcgaagct  | aagagggcag  | 540  |
| agatcatctc  | ggatcatctc | caagacgatg | agcaggtctc  | tatcatctga  | ctcgcaatca  | 600  |
| aaaagtctcg  | gttttactcc | tcctgaaaac | atgaagggtg  | atcttagcaa  | attggaaatg  | 660  |
| cctgctttac  | tgaattattg | gcgacatatt | aaccttggtg  | atgcaattcc  | aaatccatca  | 720  |
| aaggagcaac  | taattgacat | tgttcaaagg | cacttcatgt  | ctcagcaaat  | ggatgagctt  | 780  |
| caggttattg  | tggggtttgt | ccaagctgca | aagagaatga  | agaaggcttg  | caagtttcaa  | 840  |
| tccaaagaat  | ccagaaacac | tgatcttaac | tgcacagct   | aaagaaaagc  | cctgactctt  | 900  |
| aacaaatcct  | gtatgtacgg | tacatcaact | tgtttaacca  | tttgtggctt  | gctaagttaa  | 960  |
| gttcttctag  | tgatgtttgg | ctaaagggtg | gatgtttgtt  | cttctttgct  | tctgttggtt  | 1020 |
| agccaatgta  | agtaccatca | aaaaacccaa | ataactctct  | aaagctccct  | attggaaact  | 1080 |
| atcttgtctg  | atacgatctg | gagtgaccgg | tatgttggtt  | gaatgtaa    | atgtttgg    |      |

(2) INFORMATION FOR SEQ ID NO:597:

(i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 247 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS:  
(D) TOPOLOGY: linear  
(ii) MOLECULE TYPE: peptide  
(ix) FEATURE:  
(A) NAME/KEY: peptide  
(B) LOCATION: 1..247  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482509  
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:597:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Leu | Glu | Ala | Val | Asp | Ser | Ser | Gly | Val | Val | Asn | Gly | Gly | Phe | Pro |
| 1   |     |     |     | 5   |     |     |     | 10  |     |     |     |     |     | 15  |     |
| Gln | Ile | Gln | Ser | Phe | Tyr | Gly | Asp | Cys | Ser | Ser | Glu | Glu | Glu | Leu | Ser |
|     |     |     | 20  |     |     |     |     | 25  |     |     |     |     |     | 30  |     |
| Val | Leu | Pro | Arg | His | Thr | Lys | Val | Val | Val | Thr | Gly | Asn | Asn | Arg | Thr |
|     |     |     | 35  |     |     |     |     | 40  |     |     |     |     |     | 45  |     |
| Lys | Ser | Val | Leu | Val | Gly | Leu | Gln | Gly | Val | Val | Lys | Ala | Val | Gly |     |
|     |     |     | 50  |     |     |     |     | 55  |     |     |     |     |     | 60  |     |
| Leu | Gly | Gly | Trp | His | Trp | Leu | Val | Leu | Thr | Asn | Gly | Ile | Glu | Val | Lys |
|     |     |     | 65  |     |     |     |     | 70  |     |     |     |     |     | 75  |     |
| Leu | Gln | Arg | Asn | Ala | Leu | Ser | Val | Leu | Glu | Pro | Pro | Thr | Gly | Asn | Glu |
|     |     |     | 85  |     |     |     |     | 90  |     |     |     |     |     | 95  |     |
| Glu | Asp | Asp | Asp | Leu | Asp | Phe | Glu | Asn | Thr | Gln | Arg | Asn | Gly | Ser | Asp |
|     |     |     | 100 |     |     |     |     | 105 |     |     |     |     |     | 110 |     |
| Met | Ile | Val | Ser | Phe | Pro | Ala | Ser | Glu | Asp | Thr | Leu | Lys | Pro | His | Lys |
|     |     |     | 115 |     |     |     |     | 120 |     |     |     |     |     | 125 |     |
| Ser | Lys | Leu | Arg | Gly | Gln | Arg | Ser | Ser | Arg | Ser | Ser | His | Lys | Thr | Met |
|     |     |     | 130 |     |     |     |     | 135 |     |     |     |     |     | 140 |     |



Ser Arg Ser Leu Ser Ser Asp Ser Gln Ser Lys Ser Ser Gly Phe Thr  
145 150 155 160  
Pro Pro Glu Asn Met Lys Val Asp Leu Ser Lys Leu Glu Met Pro Ala  
165 170 175  
Leu Leu Asn Tyr Trp Arg His Phe Asn Leu Val Asp Ala Ile Pro Asn  
180 185 190  
Pro Ser Lys Glu Gln Leu Ile Asp Ile Val Gln Arg His Phe Met Ser  
195 200 205  
Gln Gln Met Asp Glu Leu Gln Val Ile Val Gly Phe Val Gln Ala Ala  
210 215 220  
Lys Arg Met Lys Lys Ala Cys Lys Phe Gln Ser Lys Glu Ser Arg Asn  
225 230 235 240  
Thr Asp Leu Asn Cys Ile Ser  
245

(2) INFORMATION FOR SEQ ID NO:598:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 135 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..135
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482510

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:598:

Met Ile Val Ser Phe Pro Ala Ser Glu Asp Thr Leu Lys Pro His Lys  
1 5 10 15  
Ser Lys Leu Arg Gly Gln Arg Ser Ser Arg Ser Ser His Lys Thr Met  
20 25 30  
Ser Arg Ser Leu Ser Ser Asp Ser Gln Ser Lys Ser Ser Gly Phe Thr  
35 40 45  
Pro Pro Glu Asn Met Lys Val Asp Leu Ser Lys Leu Glu Met Pro Ala  
50 55 60  
Leu Leu Asn Tyr Trp Arg His Phe Asn Leu Val Asp Ala Ile Pro Asn  
65 70 75 80  
Pro Ser Lys Glu Gln Leu Ile Asp Ile Val Gln Arg His Phe Met Ser  
85 90 95  
Gln Gln Met Asp Glu Leu Gln Val Ile Val Gly Phe Val Gln Ala Ala  
100 105 110  
Lys Arg Met Lys Lys Ala Cys Lys Phe Gln Ser Lys Glu Ser Arg Asn  
115 120 125  
Thr Asp Leu Asn Cys Ile Ser  
130 135

(2) INFORMATION FOR SEQ ID NO:599:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1323 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1323
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482514

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:599:

ccattaccta gaacatccta atcaaaaaat tgaatgttgat gataaagtct tatctgtttc 60  
aattgatgca ggaatgatga gattgggtg ctattggtgt gaaagatggt caaaaattga 120  
tgatgatggg aactgctgat gagatagtga aagctcctga gaaggccatt gtttttgcag 180

```
agaatctacc tgaagaagcg ctagccacta atctgggta cagtgtggc cttgtcaatc 240
ttggcaacac gtgttacatg aactccacgg tgcagtgtt aaaatctgtc ccagagttga 300
aatctgcatt atccaattac tcacttgctg cccgaagcaa tgatgttgac cagactttctc 360
acatgctcac agttgccaca cgtgagttat ttggtgagct tgatagaagt gtcaatgctg 420
tttcgccttc acagttcttg atggtattac gaaaaaagta tcctcagttt agtcagttgc 480
agaatggaat gcacatgcag caggatgctg aagaatgttg gacacaactg ttatacacc 540
tttctcagtc cctaaaagca ccaacttcca gcgaaggctg tgatgtgtg aaagctctat 600
ttggtgtcaa tctccagagc aggttgctt gtcaagaaag tggcgaagaa agctcagaga 660
cagaatctgt atattctcta aaatgtcata tatcacatga agtgaaccac ttgcatgaag 720
gattaaaaca tggactgaaa ggggaacttg aaaaaacatc tcctgtctct ggccgtactg 780
cactctacgt caaggagtca cttatagatt ccttgccaag gtacttgact gttcagttcg 840
tgcggttttt ctggaaaagg gagagtaatc agaaagcaaa gatcctcagg aaagtggatt 900
acccgctggg gttggatata tttgacctt gctctgagga tcttcggaag aaactggaag 960
ctcctcgcca gaaacttaga gaggaggaag gtaaaaagct tggctttcaa actagtgtca 1020
agagtggctc aaaggacagt gatgtgaaaa tgactgatgc agaggcgtct gcaaatggaa 1080
gtggagaatc atccacagta aaccacagc aaggtacttt gagccactct tagcactagt 1140
ttgaagacca agcctaaaca atgcttccac cttgtgttct ttttggatta taayccttca 1200
tgagttaatt ttggttgaac ctttggtagt atatgttgc ggattgtgca ctttctgttt 1260
tcattctctc ttccaaacta ctttattttt gcttatagat cttaatgttc tagttttgct 1320
ttt
```

(2) INFORMATION FOR SEQ ID NO:600:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 366 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..366
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482515

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:600:

```
Met Leu Met Ile Lys Ser Tyr Leu Phe Gln Leu Met Gln Asp Asp Gly
1 5 10 15
Asp Trp Ala Ala Ile Gly Val Lys Asp Gly Gln Lys Leu Met Met Met
20 25 30
Gly Thr Ala Asp Glu Ile Val Lys Ala Pro Glu Lys Ala Ile Val Phe
35 40 45
Ala Glu Asn Leu Pro Glu Glu Ala Leu Ala Thr Asn Leu Gly Tyr Ser
50 55 60
Ala Gly Leu Val Asn Leu Gly Asn Thr Cys Tyr Met Asn Ser Thr Val
65 70 75 80
Gln Cys Leu Lys Ser Val Pro Glu Leu Lys Ser Ala Leu Ser Asn Tyr
85 90 95
Ser Leu Ala Ala Arg Ser Asn Asp Val Asp Gln Thr Ser His Met Leu
100 105 110
Thr Val Ala Thr Arg Glu Leu Phe Gly Glu Leu Asp Arg Ser Val Asn
115 120 125
Ala Val Ser Pro Ser Gln Phe Trp Met Val Leu Arg Lys Lys Tyr Pro
130 135 140
Gln Phe Ser Gln Leu Gln Asn Gly Met His Met Gln Gln Asp Ala Glu
145 150 155 160
Glu Cys Trp Thr Gln Leu Leu Tyr Thr Leu Ser Gln Ser Leu Lys Ala
165 170 175
Pro Thr Ser Ser Glu Gly Ala Asp Ala Val Lys Ala Leu Phe Gly Val
180 185 190
Asn Leu Gln Ser Arg Leu His Cys Gln Glu Ser Gly Glu Glu Ser Ser
195 200 205
Glu Thr Glu Ser Val Tyr Ser Leu Lys Cys His Ile Ser His Glu Val
```

|                                                                 |                                         |     |
|-----------------------------------------------------------------|-----------------------------------------|-----|
| 210                                                             | 215                                     | 220 |
| Asn His Leu His Glu Gly                                         | Leu Lys His Gly Leu Lys Gly Glu Leu Glu |     |
| 225                                                             | 230                                     | 235 |
| Lys Thr Ser Pro Ala Leu Gly Arg Thr Ala Leu Tyr Val Lys Glu Ser |                                         | 240 |
|                                                                 | 245                                     | 250 |
| Leu Ile Asp Ser Leu Pro Arg Tyr Leu Thr Val Gln Phe Val Arg Phe |                                         | 255 |
|                                                                 | 260                                     | 265 |
| Phe Trp Lys Arg Glu Ser Asn Gln Lys Ala Lys Ile Leu Arg Lys Val |                                         | 270 |
|                                                                 | 275                                     | 280 |
| Asp Tyr Pro Leu Val Leu Asp Ile Phe Asp Leu Cys Ser Glu Asp Leu |                                         | 285 |
| 290                                                             | 295                                     | 300 |
| Arg Lys Lys Leu Glu Ala Pro Arg Gln Lys Leu Arg Glu Glu Glu Gly |                                         | 310 |
| 305                                                             | 310                                     | 315 |
| Lys Lys Leu Gly Leu Gln Thr Ser Ala Lys Ser Gly Ser Lys Asp Ser |                                         | 320 |
|                                                                 | 325                                     | 330 |
| Asp Val Lys Met Thr Asp Ala Glu Ala Ser Ala Asn Gly Ser Gly Glu |                                         | 335 |
|                                                                 | 340                                     | 345 |
| Ser Ser Thr Val Asn Pro Gln Glu Gly Thr Leu Ser His Ser         |                                         | 350 |
| 355                                                             | 360                                     | 365 |

(2) INFORMATION FOR SEQ ID NO:601:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 364 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..364
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482516

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:601:

|                                                                 |     |     |
|-----------------------------------------------------------------|-----|-----|
| Met Ile Lys Ser Tyr Leu Phe Gln Leu Met Gln Asp Asp Gly Asp Trp |     |     |
| 1                                                               | 5   | 10  |
| Ala Ala Ile Gly Val Lys Asp Gly Gln Lys Leu Met Met Met Gly Thr |     | 15  |
|                                                                 | 20  | 25  |
| Ala Asp Glu Ile Val Lys Ala Pro Glu Lys Ala Ile Val Phe Ala Glu |     | 30  |
|                                                                 | 35  | 40  |
| Asn Leu Pro Glu Glu Ala Leu Ala Thr Asn Leu Gly Tyr Ser Ala Gly |     | 45  |
| 50                                                              | 55  | 60  |
| Leu Val Asn Leu Gly Asn Thr Cys Tyr Met Asn Ser Thr Val Gln Cys |     | 65  |
| 65                                                              | 70  | 75  |
| Leu Lys Ser Val Pro Glu Leu Lys Ser Ala Leu Ser Asn Tyr Ser Leu |     | 80  |
|                                                                 | 85  | 90  |
| Ala Ala Arg Ser Asn Asp Val Asp Gln Thr Ser His Met Leu Thr Val |     | 95  |
|                                                                 | 100 | 105 |
| Ala Thr Arg Glu Leu Phe Gly Glu Leu Asp Arg Ser Val Asn Ala Val |     | 110 |
|                                                                 | 115 | 120 |
| Ser Pro Ser Gln Phe Trp Met Val Leu Arg Lys Lys Tyr Pro Gln Phe |     | 125 |
| 130                                                             | 135 | 140 |
| Ser Gln Leu Gln Asn Gly Met His Met Gln Gln Asp Ala Glu Glu Cys |     | 145 |
| 145                                                             | 150 | 155 |
| Trp Thr Gln Leu Leu Tyr Thr Leu Ser Gln Ser Leu Lys Ala Pro Thr |     | 160 |
|                                                                 | 165 | 170 |
| Ser Ser Glu Gly Ala Asp Ala Val Lys Ala Leu Phe Gly Val Asn Leu |     | 175 |
|                                                                 | 180 | 185 |
| Gln Ser Arg Leu His Cys Gln Glu Ser Gly Glu Glu Ser Ser Glu Thr |     | 190 |
|                                                                 | 195 | 200 |
| Glu Ser Val Tyr Ser Leu Lys Cys His Ile Ser His Glu Val Asn His |     | 205 |
| 210                                                             | 215 | 220 |

Leu His Glu Gly Leu Lys His Gly Leu Lys Gly Glu Leu Glu Lys Thr  
225 230 235 240  
Ser Pro Ala Leu Gly Arg Thr Ala Leu Tyr Val Lys Glu Ser Leu Ile  
245 250 255  
Asp Ser Leu Pro Arg Tyr Leu Thr Val Gln Phe Val Arg Phe Phe Trp  
260 265 270  
Lys Arg Glu Ser Asn Gln Lys Ala Lys Ile Leu Arg Lys Val Asp Tyr  
275 280 285  
Pro Leu Val Leu Asp Ile Phe Asp Leu Cys Ser Glu Asp Leu Arg Lys  
290 295 300  
Lys Leu Glu Ala Pro Arg Gln Lys Leu Arg Glu Glu Glu Gly Lys Lys  
305 310 315 320  
Leu Gly Leu Gln Thr Ser Ala Lys Ser Gly Ser Lys Asp Ser Asp Val  
325 330 335  
Lys Met Thr Asp Ala Glu Ala Ser Ala Asn Gly Ser Gly Glu Ser Ser  
340 345 350  
Thr Val Asn Pro Gln Glu Gly Thr Leu Ser His Ser  
355 360

(2) INFORMATION FOR SEQ ID NO:602:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 355 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..355
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482517

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:602:

Met Gln Asp Asp Gly Asp Trp Ala Ala Ile Gly Val Lys Asp Gly Gln  
1 5 10 15  
Lys Leu Met Met Met Gly Thr Ala Asp Glu Ile Val Lys Ala Pro Glu  
20 25 30  
Lys Ala Ile Val Phe Ala Glu Asn Leu Pro Glu Glu Ala Leu Ala Thr  
35 40 45  
Asn Leu Gly Tyr Ser Ala Gly Leu Val Asn Leu Gly Asn Thr Cys Tyr  
50 55 60  
Met Asn Ser Thr Val Gln Cys Leu Lys Ser Val Pro Glu Leu Lys Ser  
65 70 75 80  
Ala Leu Ser Asn Tyr Ser Leu Ala Ala Arg Ser Asn Asp Val Asp Gln  
85 90 95  
Thr Ser His Met Leu Thr Val Ala Thr Arg Glu Leu Phe Gly Glu Leu  
100 105 110  
Asp Arg Ser Val Asn Ala Val Ser Pro Ser Gln Phe Trp Met Val Leu  
115 120 125  
Arg Lys Lys Tyr Pro Gln Phe Ser Gln Leu Gln Asn Gly Met His Met  
130 135 140  
Gln Gln Asp Ala Glu Glu Cys Trp Thr Gln Leu Leu Tyr Thr Leu Ser  
145 150 155 160  
Gln Ser Leu Lys Ala Pro Thr Ser Ser Glu Gly Ala Asp Ala Val Lys  
165 170 175  
Ala Leu Phe Gly Val Asn Leu Gln Ser Arg Leu His Cys Gln Glu Ser  
180 185 190  
Gly Glu Glu Ser Ser Glu Thr Glu Ser Val Tyr Ser Leu Lys Cys His  
195 200 205  
Ile Ser His Glu Val Asn His Leu His Glu Gly Leu Lys His Gly Leu  
210 215 220  
Lys Gly Glu Leu Glu Lys Thr Ser Pro Ala Leu Gly Arg Thr Ala Leu

(2) INFORMATION FOR SEQ ID NO:603:

(A) LENGTH: 630 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(ix) FEATURE:

- (A) NAME/KEY: -  
(B) LOCATION: 1..630  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482525

|             |            |            |            |            |             |     |
|-------------|------------|------------|------------|------------|-------------|-----|
| aaccttcgca  | gctagatctg | gacgcttttt | actgactaga | ctcctgacaa | tatcttcato  | 60  |
| acaaatagca  | ttacattgat | gagccatgca | tcyaatcctt | ttgctctgac | cttcattttat | 120 |
| ctgcacagta  | aaatgctccg | tccaacaatt | attgcaaaag | caatggccac | agwccattct  | 180 |
| tgtcatgtga  | tcacctggta | aatcctccat | gcaaacatca | caactcatct | gtgaagactg  | 240 |
| aggaaaggaa  | gagttgccat | attgataatc | gaagacagtg | acaccagctc | cagaaaacaa  | 300 |
| rcatatmtttt | motttctmaa | caaacacagc | aaacaacttc | tccacatccc | actggtaatg  | 360 |
| aataagaaga  | gtccgtgcat | ggtgctcctt | tattgataac | aattscatca | cccttagcaa  | 420 |
| atctttctctc | tgtgctgcta | gaagcgattc | ctgagtgatg | acctgagttg | tttgawcytt  | 480 |
| tagaggacaa  | gaggctgcaa | ttcaagactc | ttcawtatca | attccatcga | aaagaatcct  | 540 |
| catcgaggga  | gtaaatggca | agcccctctc | ctccgcgcta | aaataatcat | ccatcgatca  | 600 |
| caattctqtt  | tttcgattag | gatgttctag |            |            |             |     |

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 40 amino acids  
(B) TYPE: amino acid  
(C) STRANDEDNESS:  
(D) TOPOLOGY: linear

(ix) FEATURE:

- (A) NAME/KEY: peptide  
(B) LOCATION: 1..40  
(D) OTHER INFORMATION: / Ceres Seq. ID 1482526

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Ser | His | Ala | Xaa | Asn | Pro | Phe | Ala | Leu | Thr | Phe | Ile | Tyr | Leu | His |
| 1   |     |     |     | 5   |     |     |     |     | 10  |     |     |     |     | 15  |     |
| Ser | Lys | Met | Leu | Arg | Pro | Thr | Ile | Ile | Ala | Lys | Ala | Met | Ala | Thr | Xaa |
|     |     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |
| His | Ser | Cys | His | Val | Ile | Thr | Trp |     |     |     |     |     |     |     |     |
|     |     | 35  |     |     |     |     | 40  |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:605:

- (i) SEQUENCE CHARACTERISTICS:
  - (A) LENGTH: 657 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- (ix) FEATURE:
  - (A) NAME/KEY: -
  - (B) LOCATION: 1..657
  - (D) OTHER INFORMATION: / Ceres Seq. ID 1482535
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:605:

```
attacctaga acatcctaata caaaaagtat caatggcttc cttcacctgt tcttctccat 60
cttcgatttt acctattatt gatacgagaa gtgggaattt gcgatgcaca tttcagtcctc 120
aggtttcttg tgggattcag agagatgata atggacgccg tgtttgccgg aggagaacat 180
tgacgaagaa ggacgatatg ttgcgttaca aaatgcaaag agttccattt gtggaagagc 240
aagtgaggaa gataagagra gttgggaaag taatgacaat ggacatagag cagcttcttt 300
tgagggaaga caatcggttt gaatttgtca atagcgtagc agctgaagca acagagtacg 360
tggacaagaa cagagacgaa tatggagggtt ccaaaaaagc tatctttcat gttctaagca 420
accgtgtgaa cgatctcggc ttgaccgcc ctgaggctta tgtagaagct gatccttaca 480
aaccgggtcc tggctatttg ttggagtact acacttgata tattataaca aaaagtgtca 540
atgtacttta cagcttttgt tcttgtatta ccaaaaccaa atcaatgcgt ttcacagctt 600
tgttgttttc ttggccagat ttcattttat ttatttagat ttactagatg aagacgg
```

- (2) INFORMATION FOR SEQ ID NO:606:
  - (i) SEQUENCE CHARACTERISTICS:
    - (A) LENGTH: 171 amino acids
    - (B) TYPE: amino acid
    - (C) STRANDEDNESS:
    - (D) TOPOLOGY: linear
  - (ii) MOLECULE TYPE: peptide
  - (ix) FEATURE:
    - (A) NAME/KEY: peptide
    - (B) LOCATION: 1..171
    - (D) OTHER INFORMATION: / Ceres Seq. ID 1482536
  - (xi) SEQUENCE DESCRIPTION: SEQ ID NO:606:

```
Tyr Leu Glu His Pro Asn Gln Lys Val Ser Met Ala Ser Phe Thr Cys
1 5 10 15
Ser Ser Pro Ser Ser Ile Leu Pro Ile Ile Asp Thr Arg Ser Gly Asn
20 25 30
Leu Arg Cys Thr Phe Gln Ser Gln Val Ser Cys Gly Ile Gln Arg Asp
35 40 45
Asp Asn Gly Arg Arg Val Trp Arg Arg Arg Thr Leu Thr Lys Lys Asp
50 55 60
Asp Met Leu Arg Tyr Lys Met Gln Arg Val Pro Phe Val Glu Glu Gln
65 70 75 80
Val Arg Lys Ile Arg Xaa Val Gly Lys Val Met Thr Met Asp Ile Glu
85 90 95
Gln Leu Leu Leu Arg Glu Asp Asn Arg Phe Glu Phe Val Asn Ser Val
100 105 110
Ala Ala Glu Ala Thr Glu Tyr Val Asp Lys Asn Arg Asp Glu Tyr Gly
115 120 125
Gly Ser Lys Lys Ala Ile Phe His Val Leu Ser Asn Arg Val Asn Asp
130 135 140
Leu Gly Phe Asp Arg Pro Glu Ala Tyr Val Glu Ala Asp Pro Tyr Lys
145 150 155 160
Pro Gly Pro Gly Tyr Leu Leu Glu Tyr Tyr Thr
165 170
```

- (2) INFORMATION FOR SEQ ID NO:607:
  - (i) SEQUENCE CHARACTERISTICS:
    - (A) LENGTH: 161 amino acids

- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
  - (A) NAME/KEY: peptide
  - (B) LOCATION: 1..161
  - (D) OTHER INFORMATION: / Ceres Seq. ID 1482537
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:607:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Ala | Ser | Phe | Thr | Cys | Ser | Ser | Pro | Ser | Ser | Ile | Leu | Pro | Ile | Ile |
| 1   |     |     |     | 5   |     |     |     | 10  |     |     |     |     | 15  |     |     |
| Asp | Thr | Arg | Ser | Gly | Asn | Leu | Arg | Cys | Thr | Phe | Gln | Ser | Gln | Val | Ser |
|     |     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |
| Cys | Gly | Ile | Gln | Arg | Asp | Asp | Asn | Gly | Arg | Arg | Val | Trp | Arg | Arg | Arg |
|     |     | 35  |     |     |     |     | 40  |     |     |     |     | 45  |     |     |     |
| Thr | Leu | Thr | Lys | Lys | Asp | Asp | Met | Leu | Arg | Tyr | Lys | Met | Gln | Arg | Val |
|     | 50  |     |     |     |     | 55  |     |     |     |     | 60  |     |     |     |     |
| Pro | Phe | Val | Glu | Glu | Gln | Val | Arg | Lys | Ile | Arg | Xaa | Val | Gly | Lys | Val |
| 65  |     |     |     |     | 70  |     |     |     |     | 75  |     |     |     | 80  |     |
| Met | Thr | Met | Asp | Ile | Glu | Gln | Leu | Leu | Leu | Arg | Glu | Asp | Asn | Arg | Phe |
|     |     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |
| Glu | Phe | Val | Asn | Ser | Val | Ala | Ala | Glu | Ala | Thr | Glu | Tyr | Val | Asp | Lys |
|     |     |     | 100 |     |     |     |     | 105 |     |     |     |     | 110 |     |     |
| Asn | Arg | Asp | Glu | Tyr | Gly | Gly | Ser | Lys | Lys | Ala | Ile | Phe | His | Val | Leu |
|     |     | 115 |     |     |     |     | 120 |     |     |     |     | 125 |     |     |     |
| Ser | Asn | Arg | Val | Asn | Asp | Leu | Gly | Phe | Asp | Arg | Pro | Glu | Ala | Tyr | Val |
|     | 130 |     |     |     |     | 135 |     |     |     | 140 |     |     |     |     |     |
| Glu | Ala | Asp | Pro | Tyr | Lys | Pro | Gly | Pro | Gly | Tyr | Leu | Leu | Glu | Tyr | Tyr |
| 145 |     |     |     |     | 150 |     |     |     | 155 |     |     |     |     | 160 |     |
| Thr |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:608:

- (i) SEQUENCE CHARACTERISTICS:
  - (A) LENGTH: 106 amino acids
  - (B) TYPE: amino acid
  - (C) STRANDEDNESS:
  - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
  - (A) NAME/KEY: peptide
  - (B) LOCATION: 1..106
  - (D) OTHER INFORMATION: / Ceres Seq. ID 1482538
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:608:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Leu | Arg | Tyr | Lys | Met | Gln | Arg | Val | Pro | Phe | Val | Glu | Glu | Gln | Val |
| 1   |     |     |     | 5   |     |     |     |     | 10  |     |     |     |     | 15  |     |
| Arg | Lys | Ile | Arg | Xaa | Val | Gly | Lys | Val | Met | Thr | Met | Asp | Ile | Glu | Gln |
|     |     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |
| Leu | Leu | Leu | Arg | Glu | Asp | Asn | Arg | Phe | Glu | Phe | Val | Asn | Ser | Val | Ala |
|     |     | 35  |     |     |     | 40  |     |     |     |     |     | 45  |     |     |     |
| Ala | Glu | Ala | Thr | Glu | Tyr | Val | Asp | Lys | Asn | Arg | Asp | Glu | Tyr | Gly | Gly |
|     | 50  |     |     |     |     | 55  |     |     |     |     | 60  |     |     |     |     |
| Ser | Lys | Lys | Ala | Ile | Phe | His | Val | Leu | Ser | Asn | Arg | Val | Asn | Asp | Leu |
| 65  |     |     |     | 70  |     |     |     |     | 75  |     |     |     |     | 80  |     |
| Gly | Phe | Asp | Arg | Pro | Glu | Ala | Tyr | Val | Glu | Ala | Asp | Pro | Tyr | Lys | Pro |
|     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |     |
| Gly | Pro | Gly | Tyr | Leu | Leu | Glu | Tyr | Tyr | Thr |     |     |     |     |     |     |
|     |     | 100 |     |     |     |     | 105 |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:609:

- (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 814 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (genomic)
- (ix) FEATURE:
  - (A) NAME/KEY: -
  - (B) LOCATION: 1..814
  - (D) OTHER INFORMATION: / Ceres Seq. ID 1482542

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:609:

```
agcaccggac cacacaatct tcccccaaat ctgccttcca tgcctctctt ccacggaaat 60
ctcagcacac caatcatggg acttatctct ctttactttc aatattttct cacttttctaa 120
tcctatcctt ctaattttat ttagatgtca atcattgtga ataggattat gaggctgctt 180
agttgcattg gactccaatt ggggtgaatt gctaagaaat ttcgacatgg tgcgttcata 240
taattcgact gcaacctcta caaagctgga aaacatgaat gaacatacgc ctctgtggat 300
tgcacagatc tctatctgct tctttttgga tgaacggagg gagaaagacc taggcatact 360
cagtgatccc atgaattttg tgctcctagg tacatcattt ggggctcgta cagtgtagtt 420
gtgaatctca ctaagatgcc aacgagacct tccaagaaat cagttgcata cctgcttggt 480
cgtgctccac ttctttgaat cagatggatt gcatgcagaa ttccagacac tatgctctcg 540
ttgaccagtg ctaaattaag agtcagattt tgatgaggaa gtttagcaag taacttggct 600
gaaacagcat gcttttctgt tatttgattt gcttctactg ggcactggat aagattctct 660
ggctgacctc tagttttaca tagtctctct gatagcgtgt gaccgatgta gggggtaagg 720
gatatcagaa gttttaatgc tccaacctct aattcgtcat gaggattggt gatgagttct 780
atcatggcaa agcttgcgtc ggtttctttg atcg
```

(2) INFORMATION FOR SEQ ID NO:610:

- (i) SEQUENCE CHARACTERISTICS:
  - (A) LENGTH: 52 amino acids
  - (B) TYPE: amino acid
  - (C) STRANDEDNESS:
  - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
  - (A) NAME/KEY: peptide
  - (B) LOCATION: 1..52
  - (D) OTHER INFORMATION: / Ceres Seq. ID 1482543

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:610:

```
Ala Pro Asp His Thr Ile Phe Pro Gln Ile Ser Pro Pro Ser Pro Leu
1 5 10 15
Ser Thr Glu Ile Ser Ala His Gln Ser Trp Asp Leu Ser Leu Phe Thr
 20 25 30
Phe Asn Ile Phe Ser Leu Ser Asn Pro Ile Leu Leu Ile Leu Phe Arg
 35 40 45
Cys Gln Ser Leu
50
```

(2) INFORMATION FOR SEQ ID NO:611:

- (i) SEQUENCE CHARACTERISTICS:
  - (A) LENGTH: 63 amino acids
  - (B) TYPE: amino acid
  - (C) STRANDEDNESS:
  - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: peptide
- (ix) FEATURE:
  - (A) NAME/KEY: peptide
  - (B) LOCATION: 1..63
  - (D) OTHER INFORMATION: / Ceres Seq. ID 1482544

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:611:

```
Met Val Arg Ser Tyr Asn Ser Thr Ala Thr Ser Thr Lys Leu Glu Asn
1 5 10 15
Met Asn Glu His Thr Pro Leu Trp Ile Ala Gln Ile Ser Ile Cys Phe
```



|     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|
|     | 20  |     | 25  |     | 30  |     |
| Phe | Leu | Asp | Glu | Arg | Arg | Glu |
|     | 35  |     | 40  |     | 45  |     |
| Met | Asn | Phe | Val | Leu | Leu | Gly |
|     | 50  |     | 55  |     | 60  |     |

(2) INFORMATION FOR SEQ ID NO:612:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 47 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..47
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482545

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:612:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Asn | Glu | His | Thr | Pro | Leu | Trp | Ile | Ala | Gln | Ile | Ser | Ile | Cys | Phe |
| 1   |     |     | 5   |     |     |     |     |     | 10  |     |     |     |     | 15  |     |
| Phe | Leu | Asp | Glu | Arg | Arg | Glu | Lys | Asp | Leu | Gly | Ile | Leu | Ser | Asp | Pro |
|     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |     |
| Met | Asn | Phe | Val | Leu | Leu | Gly | Thr | Ser | Phe | Gly | Ala | Arg | Thr | Val |     |
|     | 35  |     |     |     |     | 40  |     |     |     |     |     | 45  |     |     |     |

(2) INFORMATION FOR SEQ ID NO:613:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1982 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:

- (A) NAME/KEY: -
- (B) LOCATION: 1..1982
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482546

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:613:

|             |             |             |             |            |             |      |
|-------------|-------------|-------------|-------------|------------|-------------|------|
| gatccttgtc  | acaaaatgtc  | gcgtaattct  | tctactgatt  | tcagctcaat | cgctcaaatt  | 60   |
| cgagtttgcg  | tttagaaatt  | gaagttgact  | cttctgttct  | tgaatctatc | tccgatcggg  | 120  |
| gaactatctc  | tcagtaccag  | gagattgatc  | actccttcga  | cattgctctt | tgaattcgtc  | 180  |
| ctcaagggtg  | ttaatgagct  | cgtagaagct  | accagaaatg  | gcgtccatga | gctctgggtg  | 240  |
| tgaaagcctt  | cgactttgca  | tgtttgattt  | gaggagaggt  | caaactgaag | gacaagagtt  | 300  |
| agagaaaaatt | ttgttctttt  | atcctgccga  | tttagacttc  | tcgacgcagt | tatcagtgat  | 360  |
| cgggctcagt  | gaagggttta  | ttactttttt  | tagacttttc  | tctccggagg | cggcttgtga  | 420  |
| agtgatcgaa  | gcagaaagac  | attcccatgt  | tttctatgag  | gctgaacctg | atatctggat  | 480  |
| ggttatgggt  | gtggagaaaa  | ataaggagac  | aggagcgata  | tgaggatcg  | atgcattaag  | 540  |
| gagggtgctt  | aaagaagtgc  | actcactctt  | tgtgatgttt  | cacgggtcaa | ttagggcatt  | 600  |
| aatcgaaaaa  | gaaccaacag  | gagggtttac  | ccgatcacta  | ttgtaccctg | tcatacacaga | 660  |
| ttattttaagc | acattttcaa  | tatggtctct  | ctcggaagac  | tgctgctgtg | aattttttgt  | 720  |
| tggaagaaa   | cttcagctac  | caactttccg  | tgaaactttg  | agagagcggt | gaactgttca  | 780  |
| aatgcttact  | ttagcaaggg  | acactgcagt  | tgaagttcag  | tctcttggtc | aagtactaga  | 840  |
| ttcatgtgct  | gggagcttac  | gatgtcactc  | tatgatctta  | tttcaagatc | ttttggtttc  | 900  |
| aacaaccctc  | tcagctgatg  | ataccgtcga  | cttggtttaca | tttgcggtaa | tgaggttgac  | 960  |
| ctcaaaagct  | ttctcctctg  | atacgagttc  | ttgggtcatat | ctacgtaaag | ggcctgggtc  | 1020 |
| atctgaaatc  | tcttctagat  | ctaactctggc | accggttggc  | tcaattgatt | ccctacactc  | 1080 |
| aagaaacggg  | aataacatgc  | atcatgttat  | taggccacta  | caaaatgata | agtggacaaa  | 1140 |
| agggaaagat  | gggttttctaa | taaccgatat  | ttgggggtctt | gagactggcg | gctcccctga  | 1200 |
| ttctgccatc  | cctacaattt  | ggcttcagca  | gacacaagaa  | agaatgtatc | tccttgctca  | 1260 |
| tcagcataaa  | agtctcacct  | tacttcttct  | gatgcctaca  | aatgccattg | tcaatggaga  | 1320 |
| tttaagcatc  | tcagccgtga  | aacagcaagt  | tattgaagat  | gcatactga  | gaattttgaa  | 1380 |
| aattgaagag  | aatattttcaa | gagggtgggg  | cggtgagaat  | gcttaccata | ttaaggggta  | 1440 |

```
ccgttactta gtagttgata atgacacgaa agtatccaga tcttctcctt caggaaaagt 1500
aacaacactt gcaaaggagt ctctacttgc actaaacaag cttagagaag aagtggattc 1560
agaaaaaagc cgtgcaaaag gacaggagaa agacatggaa atatgcatca gagctaagaa 1620
caatgtgtgg gtgatcgccc gtgtgaccag aggcaaagag ctttacatgg ctttggagaa 1680
aggcagcgac actcttcttg ataccacaga cgctgttggg agattcagca acaggtattg 1740
cagcggagca ttcttgatgg actaagtttt cgtgttcttt cttctggttt tggaagagg 1800
gttcttctag tttcaagtac gaagtgaata gctcagaaga agtaatgagc acttctctct 1860
cagccattaa ttttgttttg tgagaaattg cagagaggaa aacgattgtg ttcttagttg 1920
gcctgtagat atgtaacaat gatattccac gttggatcag tgcaaacaaa tccttttttg 1980
tg
```

(2) INFORMATION FOR SEQ ID NO:614:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 515 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..515

(D) OTHER INFORMATION: / Ceres Seq. ID 1482547

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:614:

```
Met Ala Ser Met Ser Ser Gly Asp Glu Ser Leu Arg Leu Cys Met Phe
1 5 10 15
Asp Leu Arg Arg Gly Gln Thr Glu Gly Gln Glu Leu Glu Lys Ile Leu
20 25 30
Phe Phe Tyr Pro Ala Asp Leu Asp Phe Ser Thr Gln Leu Ser Val Ile
35 40 45
Gly Leu Ser Glu Gly Leu Ile Thr Phe Thr Arg Leu Phe Ser Pro Glu
50 55 60
Ala Ala Cys Glu Val Ile Glu Ala Glu Arg His Ser His Val Phe Tyr
65 70 75 80
Glu Ala Glu Pro Asp Ile Trp Met Val Met Val Val Glu Lys Asn Lys
85 90 95
Glu Thr Gly Ala Ile Trp Arg Ile Asp Ala Leu Arg Arg Val Leu Lys
100 105 110
Glu Val His Ser Leu Phe Val Met Phe His Gly Ser Ile Arg Ala Leu
115 120 125
Ile Glu Lys Glu Pro Thr Gly Gly Leu Thr Arg Ser Leu Leu Tyr Pro
130 135 140
Phe Ile Thr Asp Tyr Leu Ser Thr Phe Gln Ile Trp Ser Leu Ser Glu
145 150 155 160
Asp Cys Cys Cys Glu Phe Phe Val Gly Lys Lys Leu Gln Leu Pro Thr
165 170 175
Phe Arg Glu Thr Leu Arg Glu Arg Gly Thr Val Gln Met Leu Thr Leu
180 185 190
Ala Arg Asp Thr Ala Val Glu Val Gln Ser Leu Val Gln Val Leu Asp
195 200 205
Ser Cys Ala Gly Ser Leu Arg Cys His Ser Met Ile Leu Phe Gln Asp
210 215 220
Leu Leu Val Ser Thr Thr Leu Ser Ala Asp Asp Thr Val Asp Leu Phe
225 230 235 240
Thr Phe Ala Val Met Arg Leu Thr Ser Lys Ala Phe Ser Ser Asp Thr
245 250 255
Ser Ser Trp Ser Tyr Leu Arg Lys Gly Pro Gly Ser Ser Glu Ile Ser
260 265 270
Ser Arg Ser Asn Leu Ala Pro Val Gly Ser Ile Asp Ser Leu His Ser
275 280 285
Arg Asn Gly Asn Asn Met His His Val Ile Arg Pro Leu Gln Asn Asp
```

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 290 |     |     |     |     | 295 |     |     |     |     | 300 |     |     |     |     |     |
| Lys | Trp | Thr | Lys | Gly | Lys | Asp | Gly | Phe | Leu | Ile | Thr | Asp | Ile | Trp | Gly |
| 305 |     |     |     |     | 310 |     |     |     |     | 315 |     |     |     |     | 320 |
| Leu | Glu | Thr | Gly | Gly | Ser | Pro | Asp | Ser | Ala | Ile | Pro | Thr | Ile | Trp | Leu |
|     |     |     |     | 325 |     |     |     |     | 330 |     |     |     |     | 335 |     |
| Gln | Gln | Thr | Gln | Glu | Arg | Met | Tyr | Leu | Leu | Ala | Tyr | Gln | His | Lys | Ser |
|     |     |     | 340 |     |     |     |     | 345 |     |     |     |     | 350 |     |     |
| Leu | Thr | Leu | Leu | Leu | Leu | Met | Pro | Thr | Asn | Ala | Ile | Val | Asn | Gly | Asp |
|     |     | 355 |     |     |     | 360 |     |     |     |     |     | 365 |     |     |     |
| Leu | Ser | Ile | Ser | Ala | Val | Lys | Gln | Gln | Val | Ile | Glu | Asp | Ala | Ser | Leu |
| 370 |     |     |     |     |     | 375 |     |     |     |     | 380 |     |     |     |     |
| Arg | Ile | Leu | Lys | Ile | Glu | Glu | Asn | Ile | Ser | Arg | Gly | Trp | Gly | Gly | Glu |
| 385 |     |     |     |     | 390 |     |     |     |     | 395 |     |     |     |     | 400 |
| Asn | Ala | Tyr | His | Ile | Lys | Gly | Tyr | Arg | Tyr | Leu | Val | Val | Asp | Asn | Asp |
|     |     |     | 405 |     |     |     |     | 410 |     |     |     |     | 415 |     |     |
| Thr | Lys | Val | Ser | Arg | Ser | Ser | Pro | Ser | Gly | Lys | Val | Thr | Thr | Leu | Ala |
|     |     |     | 420 |     |     |     |     | 425 |     |     |     |     | 430 |     |     |
| Lys | Glu | Ser | Leu | Leu | Ala | Leu | Asn | Lys | Leu | Arg | Glu | Glu | Val | Asp | Ser |
|     |     | 435 |     |     |     |     | 440 |     |     |     |     | 445 |     |     |     |
| Glu | Lys | Ser | Arg | Ala | Lys | Gly | Gln | Glu | Lys | Asp | Met | Glu | Ile | Cys | Ile |
| 450 |     |     |     |     |     | 455 |     |     |     |     | 460 |     |     |     |     |
| Arg | Ala | Lys | Asn | Asn | Val | Trp | Val | Ile | Ala | Arg | Val | Thr | Arg | Gly | Lys |
| 465 |     |     |     |     | 470 |     |     |     |     | 475 |     |     |     |     | 480 |
| Glu | Leu | Tyr | Met | Ala | Leu | Glu | Lys | Gly | Ser | Asp | Thr | Leu | Leu | Asp | Thr |
|     |     |     | 485 |     |     |     |     | 490 |     |     |     |     |     | 495 |     |
| Thr | Asp | Ala | Val | Gly | Arg | Phe | Ser | Asn | Arg | Tyr | Cys | Ser | Gly | Ala | Phe |
|     |     |     | 500 |     |     |     |     | 505 |     |     |     |     | 510 |     |     |
| Leu | Met | Asp |     |     |     |     |     |     |     |     |     |     |     |     |     |
|     |     | 515 |     |     |     |     |     |     |     |     |     |     |     |     |     |

(2) INFORMATION FOR SEQ ID NO:615:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 512 amino acids

(B) TYPE: amino acid

(C) STRANDEDNESS:

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

(A) NAME/KEY: peptide

(B) LOCATION: 1..512

(D) OTHER INFORMATION: / Ceres Seq. ID 1482548

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:615:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Ser | Ser | Gly | Asp | Glu | Ser | Leu | Arg | Leu | Cys | Met | Phe | Asp | Leu | Arg |
| 1   |     |     |     | 5   |     |     |     |     | 10  |     |     |     |     | 15  |     |
| Arg | Gly | Gln | Thr | Glu | Gly | Gln | Glu | Leu | Glu | Lys | Ile | Leu | Phe | Phe | Tyr |
|     |     |     | 20  |     |     |     |     | 25  |     |     |     |     | 30  |     |     |
| Pro | Ala | Asp | Leu | Asp | Phe | Ser | Thr | Gln | Leu | Ser | Val | Ile | Gly | Leu | Ser |
|     |     | 35  |     |     |     |     | 40  |     |     |     |     | 45  |     |     |     |
| Glu | Gly | Leu | Ile | Thr | Phe | Thr | Arg | Leu | Phe | Ser | Pro | Glu | Ala | Ala | Cys |
| 50  |     |     |     |     |     | 55  |     |     |     |     | 60  |     |     |     |     |
| Glu | Val | Ile | Glu | Ala | Glu | Arg | His | Ser | His | Val | Phe | Tyr | Glu | Ala | Glu |
| 65  |     |     |     |     | 70  |     |     |     |     | 75  |     |     |     | 80  |     |
| Pro | Asp | Ile | Trp | Met | Val | Met | Val | Val | Glu | Lys | Asn | Lys | Glu | Thr | Gly |
|     |     |     | 85  |     |     |     |     | 90  |     |     |     |     | 95  |     |     |
| Ala | Ile | Trp | Arg | Ile | Asp | Ala | Leu | Arg | Arg | Val | Leu | Lys | Glu | Val | His |
|     |     | 100 |     |     |     |     |     | 105 |     |     |     | 110 |     |     |     |
| Ser | Leu | Phe | Val | Met | Phe | His | Gly | Ser | Ile | Arg | Ala | Leu | Ile | Glu | Lys |
|     |     | 115 |     |     |     |     | 120 |     |     |     |     | 125 |     |     |     |
| Glu | Pro | Thr | Gly | Gly | Leu | Thr | Arg | Ser | Leu | Leu | Tyr | Pro | Phe | Ile | Thr |
|     |     | 130 |     |     |     | 135 |     |     |     |     | 140 |     |     |     |     |

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Asp | Tyr | Leu | Ser | Thr | Phe | Gln | Ile | Trp | Ser | Leu | Ser | Glu | Asp | Cys | Cys | 145 | 150 | 155 | 160 |
| Cys | Glu | Phe | Phe | Val | Gly | Lys | Lys | Leu | Gln | Leu | Pro | Thr | Phe | Arg | Glu | 165 | 170 | 175 |     |
| Thr | Leu | Arg | Glu | Arg | Gly | Thr | Val | Gln | Met | Leu | Thr | Leu | Ala | Arg | Asp | 180 | 185 | 190 |     |
| Thr | Ala | Val | Glu | Val | Gln | Ser | Leu | Val | Gln | Val | Leu | Asp | Ser | Cys | Ala | 195 | 200 | 205 |     |
| Gly | Ser | Leu | Arg | Cys | His | Ser | Met | Ile | Leu | Phe | Gln | Asp | Leu | Leu | Val | 210 | 215 | 220 |     |
| Ser | Thr | Thr | Leu | Ser | Ala | Asp | Asp | Thr | Val | Asp | Leu | Phe | Thr | Phe | Ala | 225 | 230 | 235 | 240 |
| Val | Met | Arg | Leu | Thr | Ser | Lys | Ala | Phe | Ser | Ser | Asp | Thr | Ser | Ser | Trp | 245 | 250 | 255 |     |
| Ser | Tyr | Leu | Arg | Lys | Gly | Pro | Gly | Ser | Ser | Glu | Ile | Ser | Ser | Arg | Ser | 260 | 265 | 270 |     |
| Asn | Leu | Ala | Pro | Val | Gly | Ser | Ile | Asp | Ser | Leu | His | Ser | Arg | Asn | Gly | 275 | 280 | 285 |     |
| Asn | Asn | Met | His | His | Val | Ile | Arg | Pro | Leu | Gln | Asn | Asp | Lys | Trp | Thr | 290 | 295 | 300 |     |
| Lys | Gly | Lys | Asp | Gly | Phe | Leu | Ile | Thr | Asp | Ile | Trp | Gly | Leu | Glu | Thr | 305 | 310 | 315 | 320 |
| Gly | Gly | Ser | Pro | Asp | Ser | Ala | Ile | Pro | Thr | Ile | Trp | Leu | Gln | Gln | Thr | 325 | 330 | 335 |     |
| Gln | Glu | Arg | Met | Tyr | Leu | Leu | Ala | Tyr | Gln | His | Lys | Ser | Leu | Thr | Leu | 340 | 345 | 350 |     |
| Leu | Leu | Leu | Met | Pro | Thr | Asn | Ala | Ile | Val | Asn | Gly | Asp | Leu | Ser | Ile | 355 | 360 | 365 |     |
| Ser | Ala | Val | Lys | Gln | Gln | Val | Ile | Glu | Asp | Ala | Ser | Leu | Arg | Ile | Leu | 370 | 375 | 380 |     |
| Lys | Ile | Glu | Glu | Asn | Ile | Ser | Arg | Gly | Trp | Gly | Gly | Glu | Asn | Ala | Tyr | 385 | 390 | 395 | 400 |
| His | Ile | Lys | Gly | Tyr | Arg | Tyr | Leu | Val | Val | Asp | Asn | Asp | Thr | Lys | Val | 405 | 410 | 415 |     |
| Ser | Arg | Ser | Ser | Pro | Ser | Gly | Lys | Val | Thr | Thr | Leu | Ala | Lys | Glu | Ser | 420 | 425 | 430 |     |
| Leu | Leu | Ala | Leu | Asn | Lys | Leu | Arg | Glu | Glu | Val | Asp | Ser | Glu | Lys | Ser | 435 | 440 | 445 |     |
| Arg | Ala | Lys | Gly | Gln | Glu | Lys | Asp | Met | Glu | Ile | Cys | Ile | Arg | Ala | Lys | 450 | 455 | 460 |     |
| Asn | Asn | Val | Trp | Val | Ile | Ala | Arg | Val | Thr | Arg | Gly | Lys | Glu | Leu | Tyr | 465 | 470 | 475 | 480 |
| Met | Ala | Leu | Glu | Lys | Gly | Ser | Asp | Thr | Leu | Leu | Asp | Thr | Thr | Asp | Ala | 485 | 490 | 495 |     |
| Val | Gly | Arg | Phe | Ser | Asn | Arg | Tyr | Cys | Ser | Gly | Ala | Phe | Leu | Met | Asp | 500 | 505 | 510 |     |

(2) INFORMATION FOR SEQ ID NO:616:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 501 amino acids
- (B) TYPE: amino acid
- (C) STRANDEDNESS:
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: peptide

(ix) FEATURE:

- (A) NAME/KEY: peptide
- (B) LOCATION: 1..501
- (D) OTHER INFORMATION: / Ceres Seq. ID 1482549

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:616:

|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Met | Phe | Asp | Leu | Arg | Arg | Gly | Gln | Thr | Glu | Gly | Gln | Glu | Leu | Glu | Lys |
| 1   |     |     |     | 5   |     |     |     | 10  |     |     |     |     |     | 15  |     |
| Ile | Leu | Phe | Phe | Tyr | Pro | Ala | Asp | Leu | Asp | Phe | Ser | Thr | Gln | Leu | Ser |
|     |     |     | 20  |     |     |     | 25  |     |     |     |     |     | 30  |     |     |
| Val | Ile | Gly | Leu | Ser | Glu | Gly | Leu | Ile | Thr | Phe | Thr | Arg | Leu | Phe | Ser |
|     |     | 35  |     |     |     |     | 40  |     |     |     |     | 45  |     |     |     |
| Pro | Glu | Ala | Ala | Cys | Glu | Val | Ile | Glu | Ala | Glu | Arg | His | Ser | His | Val |
|     |     | 50  |     |     |     | 55  |     |     |     |     | 60  |     |     |     |     |
| Phe | Tyr | Glu | Ala | Glu | Pro | Asp | Ile | Trp | Met | Val | Met | Val | Val | Glu | Lys |
| 65  |     |     |     |     | 70  |     |     |     | 75  |     |     |     |     | 80  |     |
| Asn | Lys | Glu | Thr | Gly | Ala | Ile | Trp | Arg | Ile | Asp | Ala | Leu | Arg | Arg | Val |
|     |     |     |     | 85  |     |     |     | 90  |     |     |     |     |     | 95  |     |
| Leu | Lys | Glu | Val | His | Ser | Leu | Phe | Val | Met | Phe | His | Gly | Ser | Ile | Arg |
|     |     |     | 100 |     |     |     |     | 105 |     |     |     |     | 110 |     |     |
| Ala | Leu | Ile | Glu | Lys | Glu | Pro | Thr | Gly | Gly | Leu | Thr | Arg | Ser | Leu | Leu |
|     |     | 115 |     |     |     |     | 120 |     |     |     |     | 125 |     |     |     |
| Tyr | Pro | Phe | Ile | Thr | Asp | Tyr | Leu | Ser | Thr | Phe | Gln | Ile | Trp | Ser | Leu |
|     | 130 |     |     |     |     | 135 |     |     |     |     | 140 |     |     |     |     |
| Ser | Glu | Asp | Cys | Cys | Cys | Glu | Phe | Phe | Val | Gly | Lys | Lys | Leu | Gln | Leu |
| 145 |     |     |     |     | 150 |     |     |     |     | 155 |     |     |     | 160 |     |
| Pro | Thr | Phe | Arg | Glu | Thr | Leu | Arg | Glu | Arg | Gly | Thr | Val | Gln | Met | Leu |
|     |     |     |     | 165 |     |     |     | 170 |     |     |     |     |     | 175 |     |
| Thr | Leu | Ala | Arg | Asp | Thr | Ala | Val | Glu | Val | Gln | Ser | Leu | Val | Gln | Val |
|     |     |     | 180 |     |     |     |     | 185 |     |     |     |     | 190 |     |     |
| Leu | Asp | Ser | Cys | Ala | Gly | Ser | Leu | Arg | Cys | His | Ser | Met | Ile | Leu | Phe |
|     |     | 195 |     |     |     |     | 200 |     |     |     |     | 205 |     |     |     |
| Gln | Asp | Leu | Leu | Val | Ser | Thr | Thr | Leu | Ser | Ala | Asp | Asp | Thr | Val | Asp |
|     | 210 |     |     |     |     | 215 |     |     |     |     | 220 |     |     |     |     |
| Leu | Phe | Thr | Phe | Ala | Val | Met | Arg | Leu | Thr | Ser | Lys | Ala | Phe | Ser | Ser |
| 225 |     |     |     |     | 230 |     |     |     |     | 235 |     |     |     | 240 |     |
| Asp | Thr | Ser | Ser | Trp | Ser | Tyr | Leu | Arg | Lys | Gly | Pro | Gly | Ser | Ser | Glu |
|     |     |     |     | 245 |     |     |     |     | 250 |     |     |     |     | 255 |     |
| Ile | Ser | Ser | Arg | Ser | Asn | Leu | Ala | Pro | Val | Gly | Ser | Ile | Asp | Ser | Leu |
|     |     |     | 260 |     |     |     |     | 265 |     |     |     |     | 270 |     |     |
| His | Ser | Arg | Asn | Gly | Asn | Asn | Met | His | His | Val | Ile | Arg | Pro | Leu | Gln |
|     |     | 275 |     |     |     |     | 280 |     |     |     |     | 285 |     |     |     |
| Asn | Asp | Lys | Trp | Thr | Lys | Gly | Lys | Asp | Gly | Phe | Leu | Ile | Thr | Asp | Ile |
|     | 290 |     |     |     |     | 295 |     |     |     | 300 |     |     |     |     |     |
| Trp | Gly | Leu | Glu | Thr | Gly | Gly | Ser | Pro | Asp | Ser | Ala | Ile | Pro | Thr | Ile |
| 305 |     |     |     |     | 310 |     |     |     |     | 315 |     |     |     | 320 |     |
| Trp | Leu | Gln | Gln | Thr | Gln | Glu | Arg | Met | Tyr | Leu | Leu | Ala | Tyr | Gln | His |
|     |     |     | 325 |     |     |     |     |     | 330 |     |     |     |     | 335 |     |
| Lys | Ser | Leu | Thr | Leu | Leu | Leu | Leu | Met | Pro | Thr | Asn | Ala | Ile | Val | Asn |
|     |     |     | 340 |     |     |     |     | 345 |     |     |     |     | 350 |     |     |
| Gly | Asp | Leu | Ser | Ile | Ser | Ala | Val | Lys | Gln | Gln | Val | Ile | Glu | Asp | Ala |
|     |     | 355 |     |     |     |     | 360 |     |     |     |     | 365 |     |     |     |
| Ser | Leu | Arg | Ile | Leu | Lys | Ile | Glu | Glu | Asn | Ile | Ser | Arg | Gly | Trp | Gly |
|     | 370 |     |     |     |     | 375 |     |     |     |     | 380 |     |     |     |     |
| Gly | Glu | Asn | Ala | Tyr | His | Ile | Lys | Gly | Tyr | Arg | Tyr | Leu | Val | Val | Asp |
| 385 |     |     |     |     | 390 |     |     |     |     | 395 |     |     |     | 400 |     |
| Asn | Asp | Thr | Lys | Val | Ser | Arg | Ser | Ser | Pro | Ser | Gly | Lys | Val | Thr | Thr |
|     |     |     | 405 |     |     |     |     |     | 410 |     |     |     |     | 415 |     |
| Leu | Ala | Lys | Glu | Ser | Leu | Leu | Ala | Leu | Asn | Lys | Leu | Arg | Glu | Glu | Val |
|     |     |     | 420 |     |     |     |     | 425 |     |     |     |     | 430 |     |     |
| Asp | Ser | Glu | Lys | Ser | Arg | Ala | Lys | Gly | Gln | Glu | Lys | Asp | Met | Glu | Ile |
|     |     | 435 |     |     |     |     | 440 |     |     |     |     | 445 |     |     |     |
| Cys | Ile | Arg | Ala | Lys | Asn | Asn | Val | Trp | Val | Ile | Ala | Arg | Val | Thr | Arg |
|     | 450 |     |     |     |     | 455 |     |     |     |     |     | 460 |     |     |     |

Table 2  
Page 309

[illegible]

## POWER OF ATTORNEY

CERES, INC.  
3007 Malibu Canyon Road  
Malibu, CA 90265

I, Richard Hamilton, Chief Financial Officer of CERES, INC. of 3007 Malibu Canyon Road, Malibu, California 90265, grant Power of Attorney and authority to empower the following attorneys to act on behalf of CERES, INC. for executing Verified Statements (Declarations) Claiming Small Entity Status to be submitted to the U.S. Patent and Trademark Office in connection with the filing of provisional or regular patent applications on behalf of CERES, INC.

Raymond C. Stewart (Reg. No. 21,066)  
Joseph A. Kolasch (Reg. No. 22,463)  
Leonard R. Svensson (Reg. No. 30,330)  
Gerald M. Murphy, Jr. (Reg. No. 28,977)  
Mark J. Nuell (Reg. No. 36,623)

This Power of Attorney is to remain in full force and effect until terminated by an official of CERES, INC.

By



Richard Hamilton

Date

9/24/98

IN THE U.S. PATENT AND TRADEMARK OFFICE

I N F O R M A T I O N   S H E E T

Applicant:        Nickolai ALEXANDROV and Vyacheslav BROVER  
Appl. No.:        NEW  
Filed:             August 11, 2000  
For:               SEQUENCE-DETERMINED DNA FRAGMENTS AND  
                     CORRESPONDING POLYPEPTIDES       ENCODED  
                     THEREBY

Priority Claimed:    2750-532P        60/148,684        August 13, 1999

Send Correspondence to:

BIRCH, STEWART, KOLASCH & BIRCH, LLP    or    **CUSTOMER NO. 2292**  
P.O. Box 747  
Falls Church, VA 22040-0747  
(703) 205-8000


The above information is submitted to advise the U.S.P.T.O.  
of all relevant facts in connection with the present application.

A timely executed Declaration in accordance with 37 C.F.R.  
§ 1.64 will follow.

Respectfully submitted,

BIRCH, STEWART, KOLASCH & BIRCH, LLP

By

  
Raymond C. Stewart, #21,066

RCS/CAV  
2750-1096P

P.O. Box 747  
Falls Church, VA 22040-0747  
(703) 205-8000